Title

Epithelial zonation along the mouse and human small intestine defines five discrete metabolic domains

Journal

Nature Cell Biology, 26(2)

ISSN

1465-7392

Authors

Zwick, Rachel K
Kasparek, Petr
Palikuqi, Brisa
et al.

Publication Date

2024-02-01

DOI

10.1038/s41556-023-01337-z

Peer reviewed

HHS Public Access

Author manuscript

Nat Cell Biol. Author manuscript; available in PMC 2024 December 18.

Published in final edited form as:

Nat Cell Biol. 2024 February ; 26(2): 250–262. doi:10.1038/s41556-023-01337-z.

Epithelial zonation along the mouse and human small intestine defines five discrete metabolic domains

Rachel K. Zwick[1], **Petr Kasparek**[1,14], **Brisa Palikuqi**[1,14], **Sara Viragova**[1,14], **Laura Weichselbaum**[1,14], **Christopher S. McGinnis**[2,14], **Kara L. McKinley**[1,3], **Asoka Rathnayake**[1], **Dedeepya Vaka**[4], **Vinh Nguyen**[5,6,7,8], **Coralie Trentesaux**[1], **Efren Reyes**[1], **Alexander R. Gupta**[5], **Zev J. Gartner**[2,9,10], **Richard M. Locksley**[11,12], **James M. Gardner**[5,7], **Shalev Itzkovitz**[13], **Dario Boffelli**[4], **Ophir D. Klein**[1,4,✉]

[1]Program in Craniofacial Biology and Department of Orofacial Sciences, University of California, San Francisco, San Francisco, CA, USA.

[2]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, USA.

[3]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA.

[4]Department of Pediatrics, Cedars-Sinai Guerin Children's, Los Angeles, CA, USA.

[5]Department of Surgery, University of California San Francisco, San Francisco, CA, USA.

[6]Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, USA.

[7]Diabetes Center, University of California San Francisco, San Francisco, CA, USA.

[8]UCSF CoLabs, University of California, San Francisco, San Francisco, CA, USA.

[9]Helen Diller Family Comprehensive Cancer Center, San Francisco, CA, USA.

[10]Chan Zuckerberg BioHub and Center for Cellular Construction 94158, University of California San Francisco, San Francisco, CA, USA.

✉**Correspondence and requests for materials** should be addressed to Ophir D. Klein. ophir.klein@cshs.org.

Author contributions

R.K.Z. and O.D.K. conceived and developed the study. R.K.Z. conceived and planned experiments. D.B. conceived several computational approaches, and supervised and verified the analytical methods. R.K.Z., C.S.M., S.I., D.V. and D.B. developed the analysis strategy and performed data analysis. R.K.Z., P.K., B.P., S.V., L.W., K.L.M., A.R., V.N., C.T. and E.R. carried out experiments. A.R.G. and J.M.G. facilitated the human intestinal tissue donation. Z.J.G., R.M.L., J.M.G. and S.I. provided intellectual review of the project content. R.K.Z., D.B. and O.D.K. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Extended data is available for this paper at https://doi.org/10.1038/s41556-023-01337-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41556-023-01337-z.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this Article.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41556-023-01337-z.

[11]Department of Medicine and Department of Microbiology & Immunology, University of California San Francisco, San Francisco, CA, USA.

[12]Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, CA, USA.

[13]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel.

[14]These authors contributed equally: Petr Kasparek, Brisa Palikuqi, Sara Viragova, Laura Weichselbaum, Christopher S. McGinnis.

Abstract

A key aspect of nutrient absorption is the exquisite division of labour across the length of the small intestine, with individual nutrients taken up at different proximal:distal positions. For millennia, the small intestine was thought to comprise three segments with indefinite borders: the duodenum, jejunum and ileum. By examining the fine-scale longitudinal transcriptional patterns that span the mouse and human small intestine, we instead identified five domains of nutrient absorption that mount distinct responses to dietary changes, and three regional stem cell populations. Molecular domain identity can be detected with machine learning, which provides a systematic method to computationally identify intestinal domains in mice. We generated a predictive model of transcriptional control of domain identity and validated the roles of *Ppar-δ* and *Cdx1* in patterning lipid metabolism-associated genes. These findings represent a foundational framework for the zonation of absorption across the mammalian small intestine.

In the small intestine (SI), regional specialization optimizes digestion by enabling distinct nutrients to be absorbed sequentially at different positions. Traditionally, the SI has been separated into three loosely defined regions: the duodenum, jejunum and ileum. These segment designations, which date back to the ancient Greeks, are thought to correlate with various absorptive processes, but their anatomical boundaries are vague[1]. In addition to differences in tissue structure and cellular composition along the length of the intestinal epithelium to support specialized functions[2-4], many genes show variable spatial expression patterns[5], as most recently illustrated by single-cell RNA sequencing (scRNA-seq) comparisons of epithelial cells from the classical regions of the mouse and human SI and colon[6-11]. However, apart from the human duodenojejunal flexure, which is suspended by the ligament of Treitz, a lack of discrete landmarks to anchor these regional definitions precludes examination of the precise organization and properties of local niches within the SI. The extent to which the three classical parts explain the complexity of regional patterns in the SI, and how these patterns respond to environmental changes such as nutrient fluctuations, pathogen exposures and disease, is not clear.

By contrast, the *Drosophila* midgut divides into 10–14 compartments, of which a subset contain intestinal stem cells (ISCs) with innate regional properties[12-15] and metabolize individual nutrients[16]. It is possible that mammals too exhibit more finely grained intra-intestinal spatial differences than have been appreciated, and that ISCs program functional environments within the tissue. Indeed, regional expression of numerous genes is maintained in mouse and human intestinal organoid cultures ex vivo[17-19]. However, the molecular

programs encoded in ISCs that specify regionalized gene expression are unknown. Here we report the properties of five previously undefined epithelial regions within the mouse and human SI and establish a cellular and molecular model of their maintenance by epithelial-intrinsic mechanisms.

Results

Five enterocyte groups occupy distinct zones of the small intestine

We took an unbiased approach to define the organization of the intestine on a molecular level, asking how many functional domains, defined by distinct cellular states, are in the mammalian SI? Although previous studies assumed the presence of the duodenum, jejunum and ileum, and sampled the intestine to approximate their positions[5-11], we examined the SI without preconceptions. Our approach leveraged multiplexed scRNA-seq[20] to barcode cells collected from 30 equally sized segments spanning the entire length of the SI of both mouse and human (Fig. 1a). We used tissue from two Lgr5-GFP mice in which ISCs and their immediate transit amplifying (TA) cell progeny express green fluorescent protein (GFP), and from two human donors. We sequenced total epithelial cells ($CD45^-$, pan-epithelial $EpCAM^+$) and a similar number of progenitor cells (crypt marker $CD44^+$ in mouse and human cells, $Lgr5\text{-}GFP^+$ in mouse cells). We recovered 19,847 mouse cells and 36,588 human cells (Fig. 1a, Extended Data Figs. 1 and 2 and Supplementary Figs. 1-3), including all progenitor and specialized intestinal epithelial cell types (Fig. 1b,c and Supplementary Fig. 4), apart from $CD45^+$ tuft cells[6].

Visualization of the 30 segments in gene expression space revealed pronounced shifts in cell state along the proximal:distal axis (Fig. 1d,e). Although regionally variable genes were evident in all epithelial cell types, including secretory cells (Extended Data Fig. 3a and Supplementary Table 1), such shifts were most stark in enterocytes, with >80% of enterocyte genes in mouse and human being significantly zonated along the longitudinal axis (Methods). In the mouse, vertical zonation from the crypt/villus base boundary to the tip of the villus, previously studied only in the jejunum[21], was maintained across the proximal:distal axis (Extended Data Fig. 3b-e). These data demonstrate the impact of cell position along multiple axes on enterocyte gene expression.

We next asked whether transcriptional progression along the proximal:distal axis of the SI is continuous, or if discontinuous transitions in gene expression divide the duodenum, jejunum and ileum and/or an alternative set of regions. We computed the average expression of the 150 most regionalized genes in enterocytes, the most highly zonated epithelial cell type, from each segment and performed hierarchical clustering on the resulting data (Fig. 1f,g and Extended Data Fig. 4a, left). Remarkably, this computational approach reconstructed the anatomical order of segments in the mouse SI with almost perfect accuracy (see the segment numbers in the dendrogram in Fig. 1f, where all segments are in the correct numerical order apart from segments 14–16 and 25), reinforcing the primacy of regional position in defining enterocyte transcriptional states. We also observed essentially perfect anatomical ordering of human segments, which were grouped into pairs due to cell number variability between individual segments (see segment pair numbers in the dendrogram in Fig. 1g,

ordered accurately except for the missing pair 19–20, from which insufficient cell numbers were captured).

We also used hierarchical clustering to define the segments' higher-level organization. Specifically, the Euclidian distance between enterocyte gene expression in individual segments measured which segments had most similar expression profiles, and clustered them accordingly. The resulting hierarchical clusters (Fig. 1f,g and Extended Data Fig. 4a) revealed the order in which segments form groups at increasingly higher levels. We used the gap statistic to estimate the optimal number of enterocyte clusters[22]. With this method, gap values rise more steeply with an increasing number of well-separated clusters, and rise less steeply, or remain stable, with additional unnecessary clusters. In both mouse and human, five was the peak gap value preceding a flattening of the gap statistics (Fig. 1h,i, magenta bracket and Extended Data Fig. 4a). The boundaries of five domains were stable when using fewer than 150 genes, indicating that a five-domain superstructure is independent of the number of genes used for its identification (Extended Data Fig. 4b). Our clustering analysis revealed that mouse and human enterocytes optimally divide into five clusters of regional expression profiles, defined by the cuts of the dendrograms (Fig. 1h,i and Extended Data Fig. 4a).

We evaluated zonal enterocyte clustering based on a second metric, the Jensen–Shannon divergence (JSD), which provides a separate method to evaluate shifts in gene expression based on distances between enterocytes in UMAP space. Hierarchical clustering of JSD distances for each mouse individually provided nearly identical results to clustering based on regional gene expression (Extended Data Fig. 4c). Collectively, these data establish the positions of five intestinal domains containing transcriptionally distinct enterocytes. We have designated these regions domains A–E. On a morphological level, we observed significantly different villus lengths in domains A–D, suggesting that the overall surface area available for nutrient absorption might differ between them (Extended Data Fig. 4d).

Distinct gene signatures divide intestinal length

Given the similar number and position of domains in mouse and human, we asked whether the species might share domain-defining genes (Supplementary Tables 2 and 3). Although the regional profiles of genes such as human domain A signature gene adenosine deaminase (*ADA*) differed, we observed a correlation between many of the most highly regionalized tissue patterning and nutrient metabolism-associated genes in both species (Fig. 2a,b). We then calculated the mean scaled expression of the top 20 domain signature genes along the intestine, and plotted these domain-defining expression patterns (domain signature scores; Fig. 2c,d and Extended Data Fig. 4e). Domains A, D and E had regionally confined scores, illustrating their distinct transcriptomic signatures. Domains A and B directly overlap in the proximal-most intestine of both species, the key difference being that a small set of unique domain A-specific genes (for example, homeobox gene *Meis2*; Fig. 2b and Extended Data Fig. 4f) decline sharply, whereas genes common to both groups gradually decline over a larger area. Although domain C displayed the least zonated molecular profile of the five domains, reflected by the broad expression of genes that peak in domain C (for example, sucrase isomaltase, *Sis*), its expression pattern was clearly distinct from those in

neighbouring domains in both species. Domain E-associated transcripts (for example, the ileal fatty acid binding protein 6 *Fabp6*) emerged where domain D declined, and maintained high expression at the extreme distal end of the SI.

We then investigated the larger gene expression programs underlying the five domains using non-negative matrix factorization (NMF). NMF detects co-expressed gene modules and, unlike the signature score approach, is agnostic to putative domain boundaries. We detected modules that displayed variability across the mouse and human SI (Fig. 2e,f and Supplementary Table 4). Many of these modules contained top regional signature genes, and their expression trajectories grouped into patterns that recapitulated the signature scores. We observed two groups of components that were highly expressed at the proximal end of the intestine and declined across different breadths (as with domains A and B); components that rose and fell within the boundaries of the SI that organize into two groups—one that peaked around the centre of the intestine and one that peaked mid-way through the distal half of the intestine (roughly within the boundaries of domains C and D); and finally components that increased concurrently with the decline of domain D-associated components and did not decline (as with domain E). Thus, our NMF analysis reinforced the presence of five major patterns of regional gene expression by enterocytes across the intestine.

Detection of domain identity to delineate intestinal regions

We used multiplexed single-molecule in situ hybridization (ISH) to validate the domain assignments by probing multiple regional signature genes across full-length murine intestinal tissue (Fig. 3a, Extended Data Fig. 5 and Supplementary Fig. 5) and human tissue collected from precise positions (Fig. 3b,c and Extended Data Fig. 6). In mice, segregated localization was observed for *Meis2*, *Fabp1*, *Plb1* and *Fabp6*, markers of domains A, A/B, D and E, respectively. In human, domain A can be distinguished by human-specific domain A marker *ADA* and domain D by *PLB1*, as in mice. *SLC10A2* and *FABP6* are both expressed in domains D and E, with highest levels observed in domain E. These data support the patterns identified by scRNA-seq and highlight the transitions in regional gene expression on a tissue level.

We then sought to use the domain structure to predict domains in other datasets. We employed a machine learning approach called transfer learning[23] to train a classifier on the gene expression patterns of the domains defined by our mouse data. We then used the classifier to predict the domain identities of enterocytes from a second cohort of two mice for which we collected data from 30 segments using the same procedure as in Fig. 1a (Supplementary Fig. 6). The domain boundaries inferred from the predictions were largely consistent with the boundaries defined in our first cohort (only the boundary between domains B and C was shifted by two to three segments, Fig. 3d). These data demonstrate that domain properties can be used to predict the positions of domains across datasets.

We then used the trained classifier to predict the domain identities of cells sequenced in the original single-cell survey of the murine SI[6], in which cells were categorized as deriving from the duodenum, jejunum or ileum (Fig. 3e). Although there is no consistent method to define the traditional regions of the murine SI, based on the authors' methodologies for dividing its length[6], we estimated that the duodenum would align predominantly with

domains A and B, the jejunum with B–D, and the ileum with E and a small portion of D. We found that domain predictions for most or all cells deriving from the duodenum, jejunum and ileum aligned closely with our expectations. As with our second dataset, the model predicts fewer domain A cells and more domain C cells in the duodenal sample than expected, which may reflect minor differences in sampling strategies and is consistent with our observation that the position of the domains B–C boundary is more difficult to predict than others (Fig. 3d). Overall, the machine learning results support the presence of multiple distinct and recognizable transcriptomic signatures that align with five domains in the SI.

Domains reflect functional zones of nutrient metabolism

To evaluate whether the computationally defined domains reflect meaningful differences in intestinal function, we analysed differentially expressed genes in enterocytes from each domain associated with nutrient absorption (Fig. 4a and Supplementary Table 5). In both species, domains A and B were most strongly associated with the metabolism of fatty acids, domain C with carbohydrate metabolism, domain D with chylomicron and lipoprotein metabolism (which was also highly enriched in domain C in human) as well as amino acid transport, and domain E with cholesterol and steroid metabolism. In line with the high degree of transcriptional overlap between domains A and B (Fig. 2c-f), these domains were associated with many common processes, although in mouse, domain A was uniquely associated with iron uptake, and in both species, it displayed distinct transcripts associated with ion handling. Although domain C was largely defined by lack of expression of genes found in other domains (Fig. 1f,g), it was characterized by the highest expression of genes belonging to the carbohydrate transcriptional program[24], indicating that domain C also performs a distinct physiological role. We similarly analysed relevant NMF components (Fig. 2e,f), which provided a view not restricted by domain boundaries, of the regional span of co-expressed genes that encode nutrient metabolism proteins (Extended Data Fig. 7). For example, the formation of chylomicrons was more substantially enriched in domains C and D as above, but was detected at lower levels across domains A–D in both species. Overall, the regional patterns we identified were highly similar between the mouse and human intestine and reflect major aspects of nutrient absorption.

These functional analyses suggest that the highest levels of lipid and carbohydrate metabolism occur in distinct domains when mice are fed standard chow: fatty acid metabolism most prominently in domains A and B, phospholipid metabolism in domain D, and carbohydrate absorption broadly across the intestine but peaking in domain C. We hypothesized that enterocytes within these domains would differentially upregulate transcripts encoding the enzymes, receptors and/or binding proteins needed to absorb an increased lipid or carbohydrate dietary load. To test this prediction, we fed mice either standard chow, a high-fat/low-carbohydrate diet, or an isocaloric high-carbohydrate/low-fat diet[24]. After seven days[24-27], we sequenced single epithelial cells as in Fig. 1a from 15 equally sized segments across the intestine such that segment 1 corresponded to previously sequenced segments 1 and 2, and so on. We obtained 27,881 absorptive epithelial cells from three mice for each diet (Supplementary Fig. 7).

We applied the domain classifier (Fig. 3d) to predict the domains of cells from mice fed each diet. The resulting prediction curves (Fig. 4b) tracked the presence and position of the five domains, regardless of diet, in support of the robust nature of the domain identity, despite major dietary changes. However, domain C in mice fed a high-carbohydrate diet extended into regions normally occupied by domains B and D (green line in segments <5 and >10 in the high-carbohydrate diet; Fig. 4b). In the segments of peak domain D prediction, a similar or higher percentage of cells were classified with a domain C identity, which may suggest that enterocytes with both domain properties co-reside at this position. Enterocytes with domain C molecular and functional properties thus occupied a wider proportion of the SI, probably either in response to dietary lipid reduction or to carbohydrate augmentation.

We used NMF to examine the gene modules and the associated functions underlying this apparent shift in regional identity. Several, but not all, domain-associated modules were differentially expressed in mice fed different diets (top half of Fig. 4c and Supplementary Table 4). Module 6 was strongly associated with carbohydrate absorption, and, indeed, we observed higher levels, over a larger region, of domain C signature genes that encode components of carbohydrate digestion, including maltase-glucoamylase (*Mgam*) and *Sis*, in mice fed a high-carbohydrate diet (Fig. 4c,d).

As previously noted, multiple NMF components collectively encode domain identity (Fig. 2e), and we also observed elevated expression of other modules such as 7 following high-carbohydrate feeding. Module 9 included signatures of both domains C and D, and we observed a diet-selective response of genes within this component. Intestines from mice fed a high-fat diet upregulated domain D-associated module 9 genes as well as domains A- and B-associated module 11, which were both functionally tied with lipid metabolism (Fig. 4c). Inspection of individual components of modules 11 and 9, respectively, revealed that domain B genes that play important roles in fatty acid metabolism[28] and domain D genes in chylomicron assembly and triglyceride metabolism were most strongly enriched, especially in their respective domains (Fig. 4d). Interestingly, domain E-associated module 10 appeared completely unaffected by these dietary interventions (Fig. 4c).

Together, hierarchical clustering of gene expression in single cells identified regionalized enterocyte domains in the mouse and human intestine that we experimentally validated using multiplexed ISH. Dietary challenge experiments demonstrated unique domain responses to individual nutrients and support the functional roles of domains A/B and D in lipid metabolism and domain C in carbohydrate absorption.

Three regional ISC populations

Having established patterns of specialized gene expression in enterocytes, we asked at what stage of differentiation these patterns emerge, focusing on the murine absorptive lineage as a model. Theoretically, enterocytes could differentiate with little to no initial regional identity and take on local metabolic programs in response to microenvironmental cues. Alternatively, enterocyte fate could be pre-determined by regionalized subpopulations of stem/progenitor cells. We found that mouse ISCs displayed localized gene expression (Fig. 1d), although less markedly than enterocytes, with 46% of genes expressed by crypt cells varying along the proximal–distal axis (Methods). We again applied Euclidian (Fig. 5a) and Jensen–Shannon

(Extended Data Fig. 8a) distance metrics to calculate expression distance and perform hierarchical clustering of ISCs. Murine ISCs assembled into three regions that were well supported by the gap statistic (Fig. 5b), with boundaries that fell within two segments of each of those that delineated absorptive domains B/C and D/E. JSD also indicated three groups, albeit with slightly different boundary positions. We favoured the positions established with Euclidian distances as they draw directly from the gene expression matrix. We refer to these populations as regional ISCs 1–3.

As ISCs constitute only ~1% of the total intestinal epithelium, they have been minimally sampled in previous reports, and our progenitor enrichment strategy enabled the detection of regional ISC markers (Supplementary Table 6). For example, in addition to known proximal and distal ISC markers (for example, *Gkn3* and *Aadac* in region 1 and *Bex4* in region 3[6,29]), ISCs differentially expressed *Ttr* and *Sycn* in region 1 and *Cd177* in region 3 (Fig. 5c). In line with previous reports[29,30], we observed bacterial response genes *Defa21* and *Defa22* enriched in region 3 ISCs (Supplementary Table 6), suggesting a possible role for the regional microbiome or immune environment in shaping crypt zones.

We confirmed the spatial specificity of a subset of ISC markers using single-molecule ISH (Fig. 5d, Extended Data Fig. 8b-c and Supplementary Fig. 8). Whereas many markers were exclusively expressed by early-lineage cells (Extended Data Fig. 8d), we also noted a few shared regional markers between ISCs and later lineage cells, including 3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (*Hmgcs2*), which encodes a ketone body production enzyme. Expression of *Hmgcs2* expanded dramatically across the SI in response to a fat-free diet, as would be expected upon initiation of ketogenesis, but other regional ISC markers, such as *Gkn3* and *Bex1*, remained stable regardless of dietary lipid levels (Extended Data Fig. 8e). Although regional gene expression in mouse and human crypt cells was not as tightly correlated as for enterocytes (Fig. 5e), many transcripts such as the classic regional identity marker *Onecut2* in region 1 ISCs, and *Hoxb* genes and *Bex1* and *4* in region 3 ISCs, showed similar expression profiles.

We then used hierarchical clustering to model the point in the absorptive lineage at which these groups branch into five distinct enterocyte domains, as for Fig. 1h-i. We found three stem cell populations that give rise to three TA cell populations, which then give rise to four groups of enterocyte progenitors that ultimately specialize into five distinct enterocyte populations (Figs. 1h and 5b).

Transcriptional control of enterocyte regional identity

Given the broad zonation detected in early absorptive lineage cells (Fig. 5b), we wondered whether regionalized programs in ISCs might contribute to establishing the fate of enterocytes in each domain. In line with this possibility, previous reports[17-19] have demonstrated that regional gene expression is maintained through long-term culture of organoids, and we observed the maintenance of domain signature genes (Supplementary Table 2), including 27% of domain A genes and 30% of domain E genes, in their respective domain-specific organoid cultures (Fig. 6a, and quantitative polymerase chain reaction (qPCR) validation of select signature genes, Fig. 6b). Although mesenchymal Wnt signals drive anterior–posterior SI patterning during morphogenesis[31,32], retention

of location-specific transcript levels in vitro suggests that, in the adult organ, aspects of regional specialization are encoded within epithelial cells. Indeed, SI patterning factors *Pdx1* and *Gata4*[32-37] are expressed by epithelial cells.

To advance our understanding of the mechanisms that delineate domains, we generated a model of epithelial-intrinsic transcription factors predicted to control domain identity. We first used the gene regulatory network inference tools ChEA3[38] and SCENIC[39] to construct a model of factors most likely to control domain-specific gene expression in enterocytes (Supplementary Fig. 9 and Supplementary Tables 7 and 8). Highly ranked factors on our list included established zonation factors *Pdx1*[32,33,35] and *Gata4*[32-34,36,37], but many others were factors not previously associated with zonation (Extended Data Fig. 9a,b).

Domain E is delineated from domain D by a sharp transition in gene expression (Fig. 2b,c,e), and it appears to be disproportionately affected by several largely regionally confined gastrointestinal diseases such as ileitis and necrotizing enterocolitis. We thus focused on domain E as a test case. Factors plotted according to differentiation state (Fig. 6c, left) generally showed one of two trajectory patterns: highest expression in early lineage cells that declines as enterocytes differentiate, and expression in differentiated enterocytes or their immediate progenitors rather than early lineage cells (Extended Data Fig. 9c,d).

We first focused on the putative patterning factors expressed most highly by ISCs and TA cells. Prominent among these candidates were homeobox genes, which pattern the early gastrointestinal tract[40], but whose role in pattern maintenance during adulthood is less well understood. Caudal type homeobox 1(*Cdx1*) was expressed most highly in early-lineage cells (Fig. 6c) and specifically in region 3 ISCs and distal human ISCs (Fig. 6d). Although the importance of *Cdx2* for the structure, function and gene expression of the adult intestine is clear[41,42], the role of *Cdx1* in the adult intestine has been more challenging to determine[41,43].

To test our prediction that *Cdx1* maintains the metabolic profile of distal regions during adult homeostasis, we used two CRISPR-Cas9 gene editing strategies (Extended Data Fig. 10a-c) to delete the gene in domain E organoids, in which its expression is normally elevated relative to domain A organoids (Fig. 6e). *Cdx1* mutant organoids showed a trend towards decreased expression of the predicted target gene *Fabp6* with both strategies (Extended Data Fig. 10d). *Fabp6* is a domain E marker that is maintained in domain E organoids (Fig. 6a,b). These data support our prediction that *Cdx1* promotes the expression of the principal gene controlling long-chain fatty-acid metabolism in the distal intestine, and more broadly that regional patterning factors expressed as early as the ISC stage can control downstream aspects of nutrient processing and domain identity in enterocytes. Other patterning factors, such as *Gata4*, which is known to repress the expression of several distal genes including *Fabp6*, probably function with *Cdx1* to control domain E identity[34].

We also tested our prediction that peroxisome proliferator activator receptor delta (*Ppar-δ)*, a known regulator of fatty-acid oxidation and intestinal metabolism[44-46], controls enterocyte genes associated with lipid processing in domain E. *Ppar-δ* modulates the ISC metabolic response to diet[45,46], and although we observed expression in early-lineage cells, this

transcription factor was representative of those enriched in late lineage cells (Fig. 6c). *Ppar-δ* was expressed at slightly higher levels in domain E than in other domains in mouse and human (Fig. 6d), a pattern that was recapitulated in long-term organoid culture (Fig. 6e). We performed CRISPR-modified deletion of *Ppar-δ* in domain E organoids in the same manner as described for *Cdx1*.

Bulk RNA-seq of *Ppar-δ* mutants and controls, and qPCR validation of a subset of results, revealed differential expression of genes and enriched pathways associated with fat metabolism, including known PPAR target genes (Fig. 6f and Extended Data Fig. 10e,f). We observed decreased expression of *Fabp6* and increased domain D-associated *Plb1* levels. Interestingly, we observed the upregulation of several genes that encode fatty-acid metabolism enzymes such as ACADL, and ACOT1 and 4, which are specifically expressed in domain A in vivo during homeostasis (Fig. 6g) and are maintained in domain A organoid cultures (Fig. 6h). *Ppar-δ* loss in domain E organoids thus shifts regional organoids to a proximal lipid metabolism profile and supports our prediction that *Ppar-δ* maintains the expression signature of domain E. *Ppar-δ* works in concert with proximally enriched *Ppar-α*[46], and our results suggest that precise regional distribution of these factors may underlie PPAR-mediated patterning of lipid absorption across the intestine.

These studies indicate that epithelial-intrinsic factors that are regionally expressed by cells at multiple stages of differentiation of the absorptive lineage participate in the stable maintenance of enterocyte domain identity across the adult intestine.

Discussion

Our investigation has identified boundaries that divide the mouse and human SI into five domains of nutrient absorption (Fig. 6i). Domain A, which probably represents the duodenum based on its length and the confined expression of the classic duodenal gene *Pdx1*, contained cells from segments upstream of the ampulla of Vater, where bile and exocrine pancreatic secretions enter the intestine. Domain B overlaps with domain A in the first 6–10% of the intestine in both species, and its proximal boundary is defined by the termination of domain A-specific genes. Our analyses predict that these two domains are seeded by a common regional stem cell and represent the site of fatty acid metabolism, among other processes.

Carbohydrate absorption-associated genes of domain C are broadly expressed lengthwise, suggesting a wide range in which sugars are absorbed and metabolized. Domain C has fewer positive markers, and we speculate that the presence of an intermediate region may allow for more plasticity to respond to environmental changes. In line with this possibility, domain C is the only domain that displayed a size-wise change when mice were fed a reduced fat/increased carbohydrate diet. Further, the hierarchical clustering approach defines domain C in the second human donor more narrowly than in the first donor and in the mouse, possibly due to dietary differences.

Genes that encode ileal-specific functions, such as vitamin B12 uptake (*Cubn*) and bile salt recycling (*Slc10a2* and *Fabp6*), are enriched in domains D and E, suggesting that

these regions best approximate the ileum. Our classification of previously published data, however, suggests that domain D is probably included in studies of the murine jejunum. In both mouse and human, domain D declines as domain E increases, with only a small degree of overlap. Domain D is responsible for amino acid uptake and plasma lipoprotein processing, and is accordingly responsive to changes in dietary lipid loads. Domain E is instead predicted to metabolize steroids and cholesterol, and, remarkably, was found to be perfectly stable alongside substantial transcriptional shifts in the domain immediately adjacent in response to acute dietary change. The findings suggest that the intestinal area known as the ileum divides into two functional distinct parts. Future studies to evaluate whether domain E is innately less malleable, or whether it adapts to dietary cholesterol levels and cholesterol-lowering drugs, would be of great interest.

Further studies are also needed to dissect the response of each domain to epithelial-extrinsic factors, such as the commensal microbiome and surrounding mesenchyme. Indeed, the SI has an impressive capacity to adapt to disruptions: bowel resection leads to a shift in the expression of regional genes[47], and parasite infection remodels crypt cell identity[48] and specialized cellular distribution[49]. How the epithelial-intrinsic organization and patterning mechanisms identified here may modulate and be modulated by the enteric micro-environment is an important question for future work.

Given the radical dietary and microbiome differences between humans and laboratory mice, the similarity of domain organization between species supports the importance of an intrinsic positional system. Ex vivo maintenance of transcription factors including *Ppar-δ* and downstream target genes that define regional metabolism lends further support to the idea that domain identity is hardwired in the adult intestine, presumably on a stem cell level. The three regional ISC populations identified here express factors predicted to direct the specialization of enterocytes within the same regions, with *Cdx1* as one validated example by which *Fabp6* in enterocytes is controlled, at least in part, by a gene expressed most highly in stem cells. Several recent studies have demonstrated that metabolic programs such as ketogenesis[50], fatty acid oxidation[51] and sterol exposure[52] can influence the behaviour of ISCs and TA cells. These data add to our growing understanding of the roles of ISCs in defining local metabolic environments within the SI.

We have introduced a machine learning-based approach to identify the positions of five domains in mice. This is a systematic method to precisely track regions of the mouse intestine, and it provides a molecular classification system that future studies can utilize for consistent identification of relevant intestinal regions. Given the high variability between human samples, we did not have a sufficient number of samples to train a classifier to consistently recognize human cells. Analysis of additional subjects will strengthen our understanding of a core human domain signature, as well as, undoubtedly, further intricacies that vary between people in diverse environments.

Finally, the similarities observed between mouse and human enteric regional organization have implications for understanding the regional distribution of regionally confined gastrointestinal diseases[53-55]. We note that necrotizing enterocolitis and ileitis most commonly affect domains D and E—important sites of dietary fat response and metabolism

—raising the intriguing possibility that lipid dynamics in these positions modulate the local epithelial, immune or microbial niche with relevance to these pathologies[56]. This study provides a molecular framework that can be used to investigate the multifactorial interactions in specific cellular neighbourhoods that may predispose specific regions to disease.

Methods

Mouse and human sample information and processing for scRNA-seq

Mice.—Male and female Lgr5$^{DTR\text{-}GFP}$[57] mice were used for the scRNA-seq and RNAscope experiments in Figs. 1 and 3, and female C57BL/6J (Jackson Laboratory Strain #000664, used one week after arrival) for diet-modulation scRNA-seq experiments. Regional organoids to assess the maintenance of regional signatures were generated from adult C57BL/6J mice, and for CRISPR modulation from Lgr5^{creERT2} (ref. 58), *ROSA26*$^{LSL\text{-}Cas9\text{-}eGFP/+}$ (Jackson Laboratory strain #026175)[59] or ROSA26^{tdTomato} (Jackson Laboratory strain #007905)[60] mice (strategy 1) or Lgr5$^{DTR\text{-}GFP}$ mice (strategy 2). Mice were 8–16 weeks of age at the start of each experiment. Previously defined[24] specialized, purified high-fat/low-carbohydrate (TD.220499) and high-carbohydrate/low-fat (TD.200824) diets were purchased from Envigo and administered for seven days. All mice were housed in pairs or trios with a 7:00–19:00 light cycle with 67–74 °F temperatures and 30–70% humidity. Mice were fed ad libitum; all mice not fed specialized diets were fed PicoLab Mouse Diet 20 (5058) chow. Rodent work was carried out in accordance with approved protocols by the Institutional Animal Care and Use Committee at the University of California, San Francisco (UCSF).

Human intestinal tissue.—Human adult intestinal tissues were obtained from research-consented deceased organ donors at the time of organ acquisition for clinical transplantation through an IRB-approved research protocol with Donor Network West, the organ procurement organization for Northern California, in collaboration with the UCSF Viable Tissue Acquisition Lab (VITAL) Core. The first donor was a 44-year-old female with a body mass index (BMI) of 27 kg m^{-2} and the second donor a 30-year-old male with a BMI of 25 kg m^{-2}, both free of chronic and gastrointestinal diseases and cancer, and negative for HIV and COVID-19. Full-length intestinal tissues were collected after the clinical procurement process was completed, stored and transported in University of Wisconsin preservation media on ice, and delivered at the same time as organs for transplantation. The study and all VITAL core studies are IRB-designated as non-human subjects research, as tissues are from de-identified deceased individuals without associated personal health information.

Sample dissociation.—Mouse tissue. SI tissues were removed from the carcass and measured. The intestine from each mouse was lateralized, washed with RPMI (ThermoFisher) ‘FACS medium’ supplemented with 3% FBS, pen/strep, sodium pyruvate, MEM non-essential amino acids and L-glutamine, and cut into 30 pieces of equal length, or 15 pieces for dietary intervention studies. A single-cell dissociation of the intestinal epithelium was obtained as previously described[24]. Briefly, tissue was incubated in the supplemented RPMI medium described above with 5 mM EDTA and 10 mM dithiothreitol

(DTT) at 37 °C with 5% CO_2 for 20 min with agitation. The intestinal pieces were then triturated with a p1000 pipette, strained sequentially through 100-μm and 70-μm filters, and washed in RPMI containing 2 mM EDTA to separate the epithelial fraction.

Human tissue.—Donated SI tissues were stretched across an ice-covered trench drain and measured to be 546 cm (donor 1) and 667 cm (donor 2) in length. As with murine tissue, these lengths were divided into 30 equal segments, then 12-mm dermal punch biopsies (Acuderm) and dissection scissors were used to collect three to six biopsies as technical replicates from within the central 4-cm area in each segment. Punches were washed in DMEM/F12 (ThermoFisher) and PBS. Single epithelial cells were dissociated following previously published methods[61]. Briefly, the cells were dissociated in Ca/Mg-free HBSS (ThermoFisher) with 10 mM EDTA, pen/strep, HEPES, 2% FBS and freshly supplemented with 5 mM EDTA for 20–30 min at 37 °C with 5% CO_2 and agitation, and then for 15 min on ice. The cells were then triturated, treated sequentially with TrypLE (Gibco), DNAseI (Roche) and ACK lysis buffer as needed (ThermoFisher), and filtered through a 70-μm filter.

Sample barcoding via MULTI-seq.—Single murine and human cell suspensions from each segment were pelleted, washed, and resuspended with serum-free fluorescence-activated cell sorting (FACS) medium (as FBS and BSA prevent effective cell barcoding). MULTI-seq barcoding was performed as previously reported[20]: cells were suspended for 5 min on ice, first with an anchor/barcode solution, and then for 5 min on ice with a co-anchor solution. Following barcoding, cells from the proximal-most, middle and distal-most ten segments from mice and donor 1, and from segments with similar dissociated cell yields from donor 2, were pooled to help ensure relatively even sampling across the tissue length in subsequent steps.

FACS.—Pooled cells were stained with antibodies, all of which were diluted 1:100, against CD45 (anti-mouse, BioLegend cat. no. 103130; anti-human, BD cat. no. 564047), EpCAM (anti-mouse, BioLegend cat. no. 118214; anti-human, BioLegend cat. no. 324208) and CD44 (anti-mouse/human, BioLegend cat. no. 103026), and with DAPI. Live ($DAPI^-$) single epithelial cells ($CD45^-$, $EpCAM^+$) with the exception of $CD45^+$ tuft cells[6], and progenitors ($CD45^-$, $EpCAM^+$, $CD44^+$, Lgr5-DTR-GFP^+ mouse cells and $CD45^-$, $EpCAM^+$, $CD44^+$ human cells), were isolated using a BD FACSAria II equipped with FACSDiva software version 8 at the UCSF Parnassus Flow Cytometry Core. Boundaries between positive and negative cells were determined using unstained control samples, and control samples were stained for all but the relevant fluor. Plots were presented using FlowJo version 10 (Supplementary Fig. 1).

Single-cell barcoding, library preparation and sequencing.—Sorted total epithelial and progenitor-enriched cells from each species were pooled separately before processing in individual lanes with the 10x Genomics Chromium system. Library preparation was conducted according to the 10x Genomics standard protocol, with modifications for MULTI-seq barcode library assembly as previously described[20]. Briefly, a MULTI-seq primer was added to the complementary DNA (cDNA) amplification mix. In the first solid-phase reversible immobilisation (SPRI) bead clean-up step, MULTI-seq barcodes

in the supernatant were retained for subsequent clean-up. A PCR was also performed for MULTI-seq barcodes. Barcode libraries were analysed using a Bioanalyzer High Sensitivity DNA system and sequenced.

Gene expression and barcode cDNA libraries were pooled and sequenced using an Illumina NovaSeq 6000 machine at the UCSF Center for Advanced Technology (mouse samples and donor 2) and the Institute for Human Genetics (donor 1).

Analysis of single-cell sequencing data

Initial data processing.—All analysis steps were performed using RStudio unless otherwise noted. Mouse set 1 sequencing reads were aligned using CellRanger version 3.0.1 (10x Genomics) to the mouse mm10-3.0.0 reference (10x Genomics). Sequencing reads for donor 1 were aligned using kallisto-bustools v0.46.2[62] to the human GRCh38.95 reference. Sequencing reads for donor 2, mouse set 2 and the mouse diet experiment were aligned using CellRanger version 7.0.0 (10x Genomics) to the same respective references.

Raw gene expression count matrices were filtered using DropletUtils[63] to identify real cells. Demultiplexing and removal of predicted doublets and unclassified cells was done with the deMULTIplex R package[20] for mouse set 1 scRNA-seq data; with the hashedDrops function of DropletUtils for donor 1 scRNA-seq data; and with a combination of the hashedDrops function of DropletUtils and deMULTIplex2[64] for the donor 2 scRNA-seq, mouse set 2 and mouse diet data. Finally, identified cells were filtered according to the number of unique molecular identifiers per cell, the number of genes per cell and the percentage of mitochondrial gene reads per cell (Extended Data Figs. 1 and 2 and Supplementary Figs. 3, 6 and 7).

After performing sample demultiplexing on murine set 1 and donor 1 scRNA-seq data, we addressed two experimental issues computationally. First, in the murine scRNA-seq data, we noted that identical MULTI-seq sample barcodes were inadvertently applied to cells derived from segments 9–16 in the two mice sampled, as evidenced by the mix of male and female sex-linked genes in cells assigned to ‘mouse A’, and a complete lack of cells in the same regions of cells assigned to ‘mouse B’ (Supplementary Fig. 2). To distinguish between individual mouse samples, we used scPred[65] to train a classifier that assigns cells from all segments to male, female or unassigned status, and associated them to the appropriate segment position in mouse A or B accordingly (Supplementary Fig. 2b,c). Second, in the donor 1 scRNA-seq data, we noted that human cells associated with the MULTI-seq barcode for segment 30 were not recovered, which may be due to inefficient barcode labelling or sequestering of the barcode by dead cells or highly viscous mucus content in the distal-most portion of the human intestine during cell dissociation. All analysis of human data from donor 1 was therefore performed on segments 1–29, as displayed in the relevant figures.

Mouse set 1 and donor 1 data were processed in Seurat V3[66]. Donor 2, mouse set 2 and the mouse diet experiment were processed in Seurat V4 (ref. 67). For mouse sets 1 and 2, total epithelial and progenitor-enriched samples were processed with the SCTransform function[68] with 3,000 features requested, with regression of differences in cell-cycle state among cells, the level of expression of mitochondrial genes and of a set of sex-specific

genes (*Xist*, *Tsix*, *Ddx3y*, *Eif2s3y*), followed by integration with Seurat's IntegrateData function. Because the focus of mouse set 2 was on enterocytes, we did not integrate or further process cells from the progenitor-enriched fraction. The mouse diet samples were processed in the same way except for the regression of the expression of sex genes, as all the mice in this dataset were females. Donor 1 total epithelial and progenitor-enriched samples were processed with the SCTransform function with 3,000 features requested, with regression of the level of expression of mitochondrial genes, followed by integration with the fastMNN function. fastMNN integration was applied to the human scRNA-seq data because it was the most effective procedure to correct batch effects between total epithelial and progenitor-enriched samples. Donor 2 total epithelial and progenitor-enriched samples were merged and processed with the SCTransform function with 3,000 features requested, with regression of the level of expression of mitochondrial genes. Data from donor 2 did not require integration.

We performed data-dimensionality reduction using principal component analysis in Seurat for all datasets except donor 1, for which the MNN components identified with fastMNN integration were used as low-dimension components. The number of principal components used was determined for each sample by inspection of the sample's elbow plot. The following top components were used: mouse set 1, 50; mouse set 2, 32; mouse diet, 30; donor 2, 36. Finally, for donor 1 we used the first 50 MNN components. We also tested the stability of the downstream results (number of identified clusters, shape of the UMAP) to different choices of number of top principal components. Following dimensionality reduction, the nearest-neighbour graph was calculated with the Seurat function FindNeighbors with the default argument k.param=20. We then identified clusters using the Seurat function FindClusters with default resolution (resolution=0.8), except for donor 1, for which we used a resolution of 0.55.

We classified the cell type identities of cells from mouse set 1 using Seurat to project previously reported reference cell type annotations for the murine intestinal epithelium[6] onto the present data (Extended Data Fig. 1 and Supplementary Fig. 4). Cell type annotation was refined by intersecting the transferred annotations and the clusters identified using Seurat, and resolving ambiguities using the following algorithm: (1) clusters in which most cells had the same transferred annotation (this was the case for all clusters except cluster 15): cells annotated with the majority annotation were retained, cells without the majority annotation were annotated as 'unknown' and not included in the analysis of regionality; (2) cluster containing cells with two annotations transferred at high frequency: one cluster (cluster 15) contained mostly cells annotated as either 'transit amplifying' or 'enterocyte'. Cells annotated as one of these two types were retained, and all other cells were annotated as 'unknown' and not included in the analysis of regionality. Overall, cells of unknown identity constituted 7.6% of the total number of cell post-quality control in the mouse dataset, but did not group into a single cluster.

All other single cells were annotated by assigning cell-type identities based on marker gene expression[6,7] (Supplementary Figs. 4, 6 and 7). Clusters showing moderate expression of both cycling_g2m and enterocyte genes were annotated as 'enterocyte progenitors'; this annotation was also supported by the spatial observation that clusters annotated

as enterocyte progenitors were found between TA cells and enterocytes in the UMAP visualization of the cells of the human dataset. Outlier cells that could not be annotated using existing marker genes (<2% of cells in either donor) were removed.

Seurat was used throughout our analysis for the generation of violin plots, dot plots, ridge plots and marker lists.

Villus zonation scoring.—MATLAB version 2018b was used to annotate the enterocytes according to their position along the crypt:villus axis using our previously published strategy[21]. Villus zonation scores draw from the summed expression of landmark genes[21] and represent the ratio of the summed expression of the top landmark genes (tLM) and the summed expression of the bottom (bLM) and tLM genes (Extended Data Fig. 3). tLM and bLM were chosen based on the single cell-reconstructed zonation profiles as in ref. 21, as genes with a sum-normalized expression above 10^{-3} in at least one of the six villus zones and a centre of mass above 3.5 for tLM or below 2.5 for bLM. The centre of mass is average zone weighted by the expression of the respective gene[21]. An equal number of cells within the enterocyte clusters were assigned to each of six crypt:villus zones, zones 1–6 (Extended Data Fig. 3).

Calculation of percent regionalization and gene expression distance across segments.—The Kruskal–Wallis test was used to calculate the percent of regional zonation among genes with mean sum-normalized expression above 5×10^{-6}. This analysis was only possible for cell types with >40 cells per domain. The q values were produced using the Benjamini–Hochberg procedure for multiple hypotheses correction. The false discovery rate was set at $q < 0.05$. The centres of mass for all enterocyte-expressed genes (Fig. 2a), crypt-expressed genes (Fig. 5e) and gene markers of specific secretory cell types (Supplementary Table 1) were calculated across even fifths of the length of the intestine. For mouse–human correlations, we compared the segment centres of mass using a mouse–human orthology table based on Ensembl (version 109)[69] using the BioMart data-mining tool. Genes with a sum-normalized expression above 10^{-5} in at least one of the five segments are shown in the scatterplots in Figs. 2a and 5e. Genes with the highest and lowest segmental centres of mass (reflecting the proximal and distal-most expressed genes) and those with median centres of mass and highest Euclidean distance between the segmental profiles normalized to their maximum (reflecting the centre-most expressed genes) were labelled, and coloured according to domain identity (Supplementary Table 2) if applicable.

Heatmaps were generated using pheatmap[70], with the average normalized expression of the 150 genes most highly upregulated per segment in enterocytes (defined as the combination of cells annotated as differentiated or mature enterocytes; Fig. 1f,g), or the top 100 marker genes per segment in ISCs (Fig. 5a). Because the cell number per segment is variable in the human dataset, segments were grouped into pairs for this analysis. Heatmaps visualize data from a matrix in which each cell contains the average expression of a marker gene in each segment. Segments and genes were clustered based on the Euclidean distance between cells in the matrix. The optimal number of clusters was identified by computing the gap statistic using the clusGap function of the R package cluster (version 2.1.4) using default parameters. We also confirmed that domain divisions were stable when alternative numbers

of top upregulated genes were used (Extended Data Fig. 4b, displaying 75–100 upregulated genes per segment).

To evaluate the domain assignments with a different approach, we calculated the JSD[71,72] for enterocytes and ISCs on the mouse dataset (Extended Data Figs. 4c and 8a). To calculate the JSD, we assigned a centre of mass to each segment by bivariate kernel density estimation and calculated the pairwise JSD between the resulting vectors. For enterocytes, JSD was calculated for each mouse individually. Mouse 2, which contains fewer cells and has greater cell-number-per-segment variability than mouse 1, had slightly weaker segment ordering (note the positions of segments 19–20) than mouse 1, but mis-ordering was confined to domains and did not ultimately affect our interpretation of appropriate boundary divisions.

The domain-defining signature score (Fig. 2c,d) is a *z*-score metric representing the mean expression of the 20 most differentially expressed genes in a given absorption domain. The signature scores were computed from scaled and centred gene expression data following SCTransfom in Seurat.

NMF analysis.—We performed NMF analysis using the cNMF package version 1.4 in R[73]. We used the raw count matrices for a given subset of cells as input to cNMF, and ran cNMF with default parameters. For visualization of the results, we selected the 250 genes with the strongest contribution to a component and used the Seurat function AverageExpression to compute the averaged expression of the selected genes.

Prediction of intestinal domain locations using transfer learning.—We performed the computational transfer of domain labels from mouse datasets with known domain assignment (training datasets) to datasets with unknown domain position (test dataset) by transfer learning using the cFIT package version 0.0.0.90 in R[23]. We used the raw count matrices for enterocytes and mature enterocytes as input to cFIT. All cells (both the training and test sets) were labelled according to their experimental batch. Cells from the training sets were also labelled according to their previously assigned domains. cFIT was run with default parameters and requesting 15 factors of the common factor matrix (shared across training and test datasets), using the datasets provided in Supplementary Table 9.

Functional pathway analysis.—Pathways enriched in each mouse and human absorption domain (adjusted $P < 0.02$; Fig. 4a and Supplementary Table 5) or regionally variable NMF component (adjusted $P < 0.04$; Extended Data Fig. 7) were identified using the ReactomePA enrichPathway tool and compared using the clusterProfiler package[74]. Selected pathways associated with nutrient metabolism are shown. Pathways were edited to remove redundancy and plotted with ggplot2.

Evaluation of transcriptional control of domain identity.—We first used ChIP-X Enrichment Analysis 3 (ChEA3)[38] to identify the transcription factors predicted to control genes differentially expressed in enterocytes from each absorption domain. We repeated this analysis for enterocytes, TA cells and ISCs, such that we might evaluate which transcription factors expressed by each of these cell types is predicted to control domain-specific expression in enterocytes. Transcription factor enrichment results generated with this

approach (Supplementary Table 7) are ranked according to several types of data, including transcription factor–gene association in RNA-seq and ChIP-seq datasets, and co-occurrences in submitted gene lists. We also used SCENIC[39,75] to infer gene regulatory networks based on co-expression and motif analysis of transcription factors and targets, which were then analysed in individual differentiated and mature enterocytes (Supplementary Table 8).

To evaluate the expression of each transcription factor along stages of absorptive cell differentiation, from ISC to enterocyte, we used Slingshot[76] to infer the differentiation pseudotime for all absorptive cells and order the cells accordingly, allowing us to then plot the expression of transcription factors across differentiation states (Fig. 6c and Extended Data Fig. 9c,d).

Transcription factors were evaluated according to their predictive rank in ChEA3, convergent identification in ChEA3 and SCENIC analyses, and regional expression across domains (Supplementary Fig. 9). We grouped the transcription factors according to the highest expression at early (ISC/TA cell) or late (enterocyte precursor or later) stages of the absorptive lineage (Extended Data Fig. 9c,d).

Visualization of regional marker transcripts

Full-length murine SI tissue or transverse cross-sections of human intestines from the indicated domains were immersed in 4% paraform-aldehyde for 24–48 h at room temperature and ethanol for 24 h at 4° C. Murine SIs were coiled into a ‘Swiss roll’ from an outer proximal tip to an inner distal tip. All tissue underwent standard dehydration and paraffin embedding.

The RNAscope Multiplex Fluorescent V2 Assay (Advanced Cell Diagnostics) was used according to the manufacturer’s instructions to probe for transcripts of interest. Entire Swiss rolls were captured with a Leica DMi8 microscope equipped with LAS X Software and an automated stage, allowing for tilescan imaging of frames at ×20 magnification. Three to five individual images were acquired per region from each donor. Regional patterns of selected individual marker transcripts were confirmed on at least three mice each and in three to four donors per domain, including the two donors sequenced in this study. Images of individual murine crypts and crypt-villus units were also captured using a Zeiss LSM900 confocal microscope.

For morphometric analysis of villus height (Extended Data Fig. 4d), the lengths of tilescanned Swiss rolls were tracked using a custom macro for Fiji[77], allowing assignment of the precise positions of 30 equal segments. Villus base-to-tip distances were measured for three to five villi in each segment, for each of four mice. One-way analysis of variance (ANOVA), followed by Tukey’s multiple comparisons test for villus heights across all segments in each domain, was performed using Prism software (GraphPad Prism version 8 for MacOS).

Human tissue images were analysed using a custom script in QuPath software[78]. Briefly, nuclei detection was performed using StarDist2D, and cell segmentation was performed with the cell expansion variable set to 10 μm. The mean fluorescence value for each cell

was plotted (Fig. 3c), and one-way ANOVA was performed in Prism to compare the mean fluorescence in each donor by domain (Extended Data Fig. 6b).

Investigation and genetic perturbation of regional organoids Generation and qPCR evaluation of regional organoids.

Intestinal crypts were isolated from domains A–E of fresh intestinal tissue using methods as previously described[79].

For the evaluation of gene expression with qPCR or mRNA-seq, organoids that had been cultured for at least one month (5–13 weeks), and 5–6 days after passaging, were washed with PBS and resuspended in TRI reagent containing 1% 2-mercaptoethanol. RNA was extracted using a Direct-zol RNA Miniprep Plus kit (Zymo Research) and cDNA was reverse-transcribed with a high capacity cDNA reverse transcription kit (Applied Biosystems) according to the manufacturer's instructions. qPCR using the primers listed in Supplementary Table 8 was performed using a C1000 Touch Thermal Cycler (Biorad).

CRISPR-mediated gene disruption.—Two single-guide RNAs (sgRNAs) were designed for each target using the Benchling CRISPR Guide RNA Design tool (https://www.benchling.com/crispr/). Following previously described methods[80] and using BstXI (Thermo Fast Digest, cat. no. FD1024) and BlpI (Thermo Fast Digest isoschizomer Bpu1102I, cat. no. FD0094) restriction enzymes, we inserted a sgRNA into the pU6sgRNA-EF1alpha-puro-T2A-BFP single cassette vector, which expresses the mouse U6 (mU6) promoter and constant region 1 (cr1)[81], and the second sgRNA into pMJ117, which expresses the modified human U6 (hU6) promoter and cr3[82]. The sgRNA sequences, and the primers used for subsequent PCR amplification (Q5 Hot Start High-Fidelity 2X Master Mix, NEB) of sgRNA expression cassettes, are provided in Supplementary Table 10. pU6sgRNA-EF1alpha-puro-T2A-BFP was then digested with XhoI and XbaI (NE Biolabs) and gel-purified along with the PCR fragments. The sgRNAs were then incorporated into the pU6sgRNA-EF1alpha-puro-T2A-BFP backbone using NEBuilder HiFi DNA Assembly Master Mix (NE Biolabs) according to the manufacturer's instructions. Lentivirus was produced from the resulting dual sgRNA constructs by UCSF Viracore. Virus was concentrated using Lenti-X concentrator (Takara Biosciences).

To increase the efficiency of CRISPR mutagenesis, we also used a second strategy based on simultaneous delivery of Cas9 and sgRNA by lentiviral vectors. Using Esp3I restriction enzyme (New England Biolabs, cat. no. R0734S), we inserted each sgRNA into lentiCRISR v2 (Addgene 52961), which allowed simultaneous expression of gRNA driven by the U6 promoter and Cas9/PuroR driven by EF1alpha. Cloning was performed as described in ref. 83, and successful insertion of the sgRNA sequence was validated by Sanger Sequencing using primer 5′-GCACCGACTCGGTGCCAC-3′. The sgRNA sequences are provided in Supplementary Table 10. Lentivirus was produced from the resulting vectors as described in ref. 83.

Lentiviral transduction of adult, regional organoids for all experiments was performed as described in ref. 84. Briefly, intestinal organoids were grown for at least four days before infection in 'ENRWNTNIC' (50% growth medium/50% Wnt-cultured medium and 10 mM

nicotinamide), supplemented with 10 μM Y-27632 and 2.5 μM CHIR 99021 to induce spheroid formation. Stem cell-enriched spheroids were broken into single cells for the addition of viral mix containing 8 μg ml^{-1} polybrene, followed by a 1-h spinoculation and a 6-h incubation at 37°. Infected cells were then plated in Matrigel. Puromycin selection was performed 72 h after recovery. Spheroids were converted to organoids over the course of ~7 days by gradual transition of ENRWNTNIC to ENR medium.

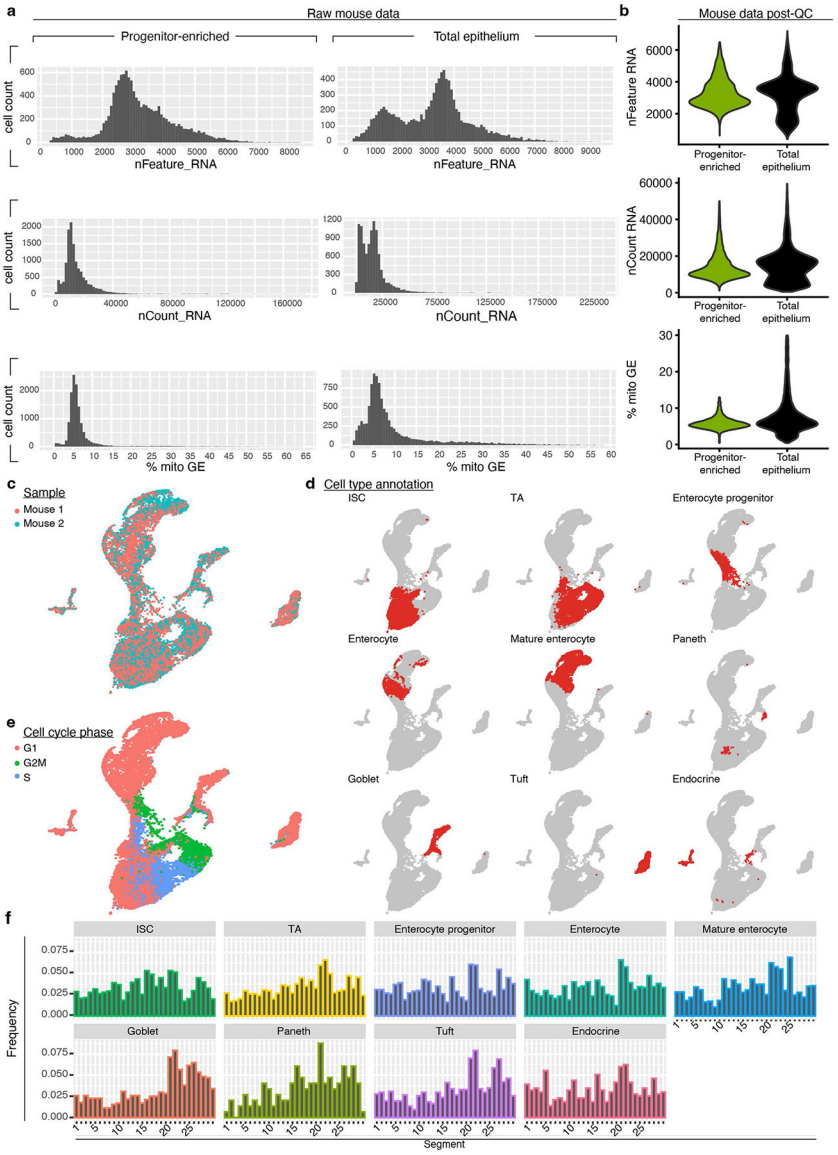
Infected organoids were expanded and, for strategy 1, treated with 4-hydroxytamoxifen to induce Cre recombinase-dependent expression of Cas9 endonuclease and enhanced green fluorescent protein (EGFP). From these cultures, organoids were passaged at a low density (strategy 2), or small numbers (1–100) of single, BFP^+ (transduced), GFP^+ (tamoxifen-induced) cells were sorted into individual, Matrigel-coated wells of a 96-well plate (strategy 1), in both cases allowing for precise manual isolation of individual organoids. After ~10 days of growth, single mature organoids were collected and used for clonal expansion. To confirm genetic disruption, genomic DNA was isolated (Lysis and Neutralization Solutions for Blood, Sigma), genotyped with PCR, and the mutant alleles were sequenced (primers, Supplementary Table 10). Clones carrying the wild-type alleles were excluded and only the clones with deleterious alleles were used for the downstream analyses.

mRNA-seq of regional organoids.—RNA was collected from confirmed mutant organoid clones, transduced organoids uninduced by 4-hydroxytamoxifen (OHT), and untreated organoids as described above for qPCR evaluation. All organoid lines were cultured for five to six days post-passaging to ensure consistent and complete differentiation status across samples. RNA sample quality control, mRNA-seq library preparation and mRNA-seq (Illumina, PE150, 20 million paired reads) was performed by Novogene.

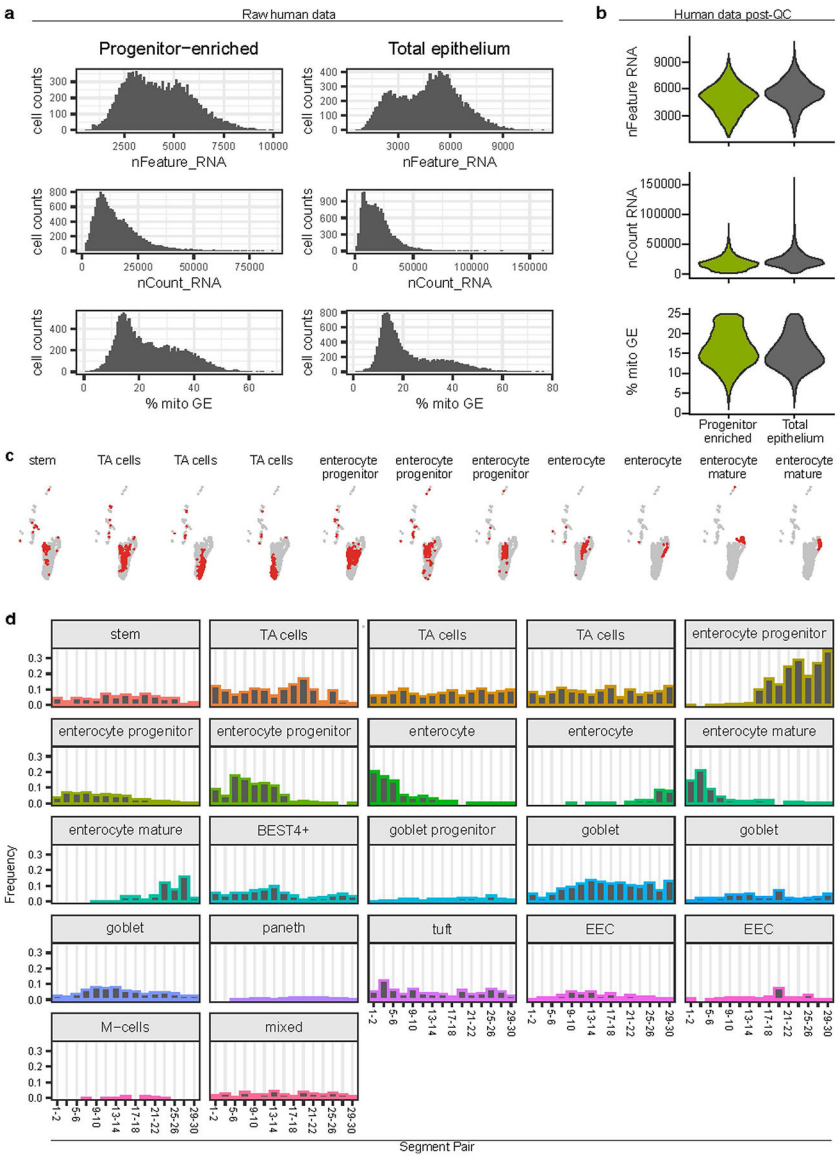
Genome indexing and quantification of transcript abundances by pseudoalignment were performed using Kallisto version 0.46.0 (ref. 85). Non-expressed genes were filtered by retaining genes with more than five reads in at least four samples. RUVseq[86] was used to control for ‘unwanted variation’ between samples. Differentially expressed genes in mutant organoids compared to untreated organoids were identified using EdgeR. Because mutant organoids were assayed without replication, data dispersion was estimated from all but the 5,000 most variable genes in the entire dataset.

Statistics and reproducibility.—No statistical method was used to predetermine sample size, but our sample sizes are similar to or greater than those reported in previous publications[6]. Only data from cells deemed low quality using the procedures described in the Initial data processing subsection of the Analysis of single cell sequencing data section of the Methods were excluded from the analyses. Administration of specialized diets was randomized. Data distribution was assumed to be normal, but this was not formally tested. The investigators were not blinded to allocation during the experiments and outcome assessment.

Extended Data

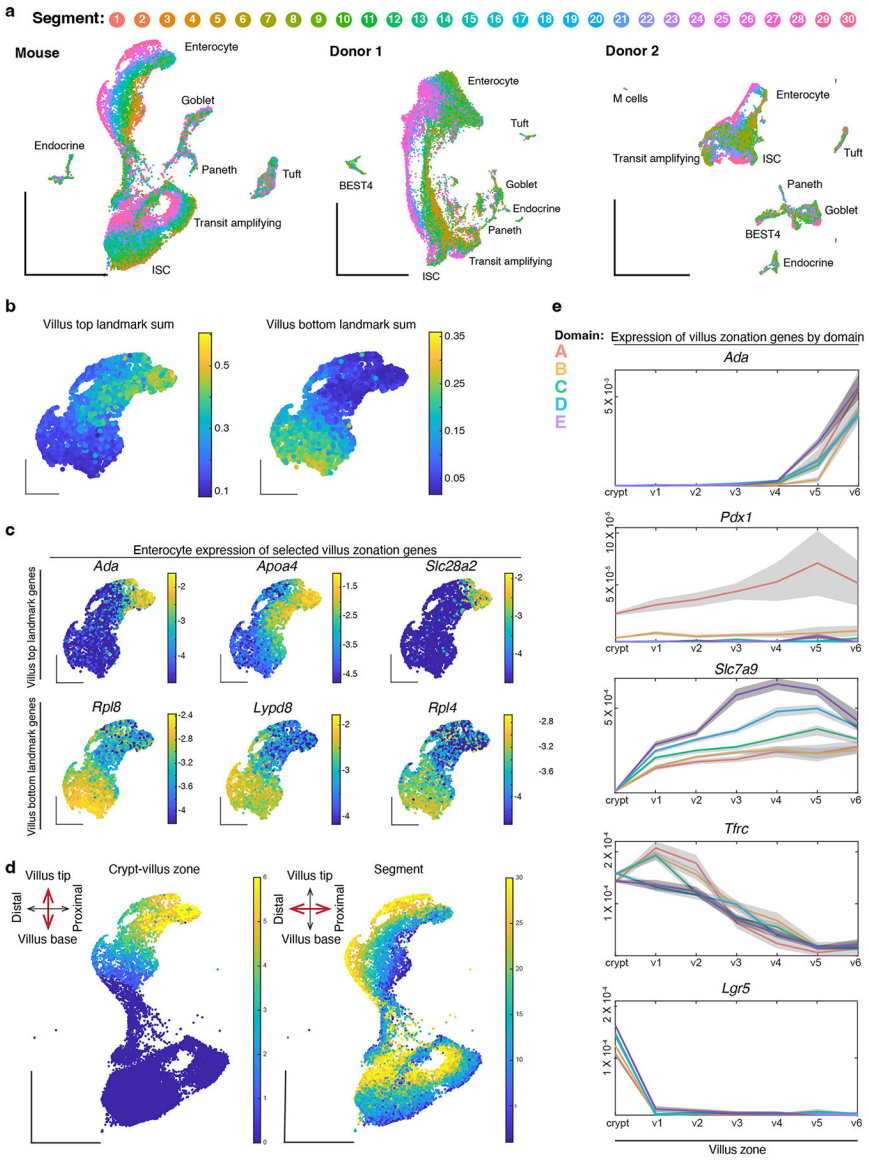


Extended Data Fig. 1 |. Quality control and initial processing of mouse scRNA-seq data. **a,b** Quality control metrics of data, including number of genes detected (‘nFeature_RNA’), number of unique molecular identifiers detected (‘nCount_RNA’), and percent mitochondrial reads (‘% mito GE) before (**a**) and after (**b**) processing data. **c-e** UMAP of total murine epithelial cells sequenced post-QC, coloured according to mouse identity (**c**), cell type annotation (**d**), or cell cycle phase (**e**). **f** Frequency of epithelial cells of indicated subtype by segment. QC, quality control, mito, mitochondrial; GE, gene expression; ISC, intestinal stem cell; TA, transit amplifying; G1, growth 1; G2M, growth 2 mitosis; S, synthesis.



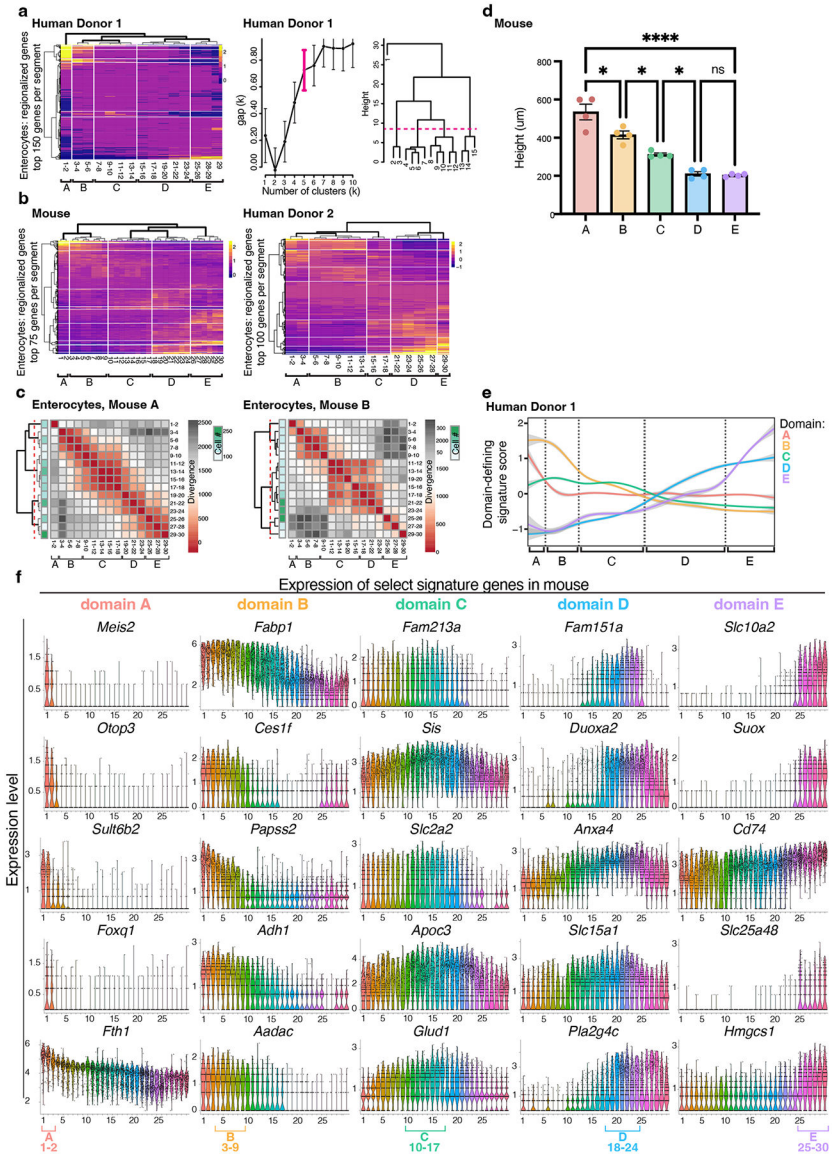
Extended Data Fig. 2 |. Quality control and initial processing of human scRNA-seq data from human subject 2.

a,b Quality control metrics of data, including number of genes detected (‘nFeature_RNA’), number of unique molecular identifiers detected (‘nCount_RNA’), and percent mitochondrial reads (‘% mito GE’) before (**a**) and after (**b**) processing data. **c** UMAP of total human cells sequenced post-QC, highlighting cell type annotation. **d** Frequency of cells of all epithelial subtypes by segment pair. QC, quality control; mito, mitochondrial; GE, gene expression; ISC, intestinal stem cell; TA, transit amplifying.



Extended Data Fig. 3 |. Zonation across multiple axes of the small intestine.

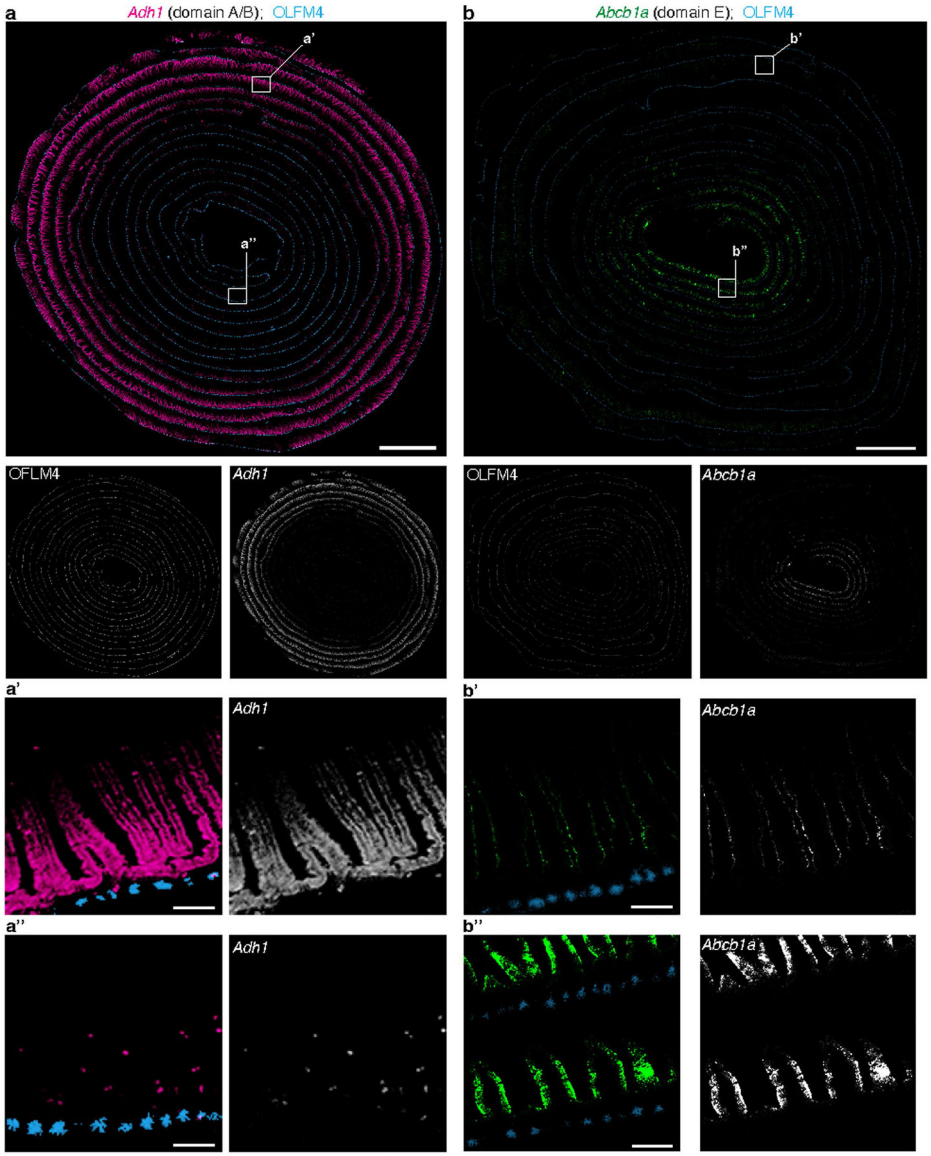
a UMAP of absorptive lineage cells coloured by segment number along the proximal to distal axis in mouse and human donors. Major epithelial cell types are labeled. **b-e** Villus zonation across murine enterocytes. **b** UMAP plots coloured according to summed expression of previously reported[21] landmarks of the villus tip (left) or base of villus (right). An equal number of enterocytes were assigned to each of 6 crypt:villus zones, zones 1–6. **c** UMAP plots coloured according to the expression of select top and bottom villus markers. **d** UMAP plots coloured according to villus zonation scores (left) compared to segment positions (right). Villus zonation scores represent the ratio of the summed expression of bottom and top landmark genes. **e** Expression of select villus zonation markers across crypt:villus zones. Center lines represent zone mean, and are coloured by domain with surrounding grey standard error bands. M-, microfold.



Extended Data Fig. 4 |. Stability and features of five domains across the mouse and human small intestine.

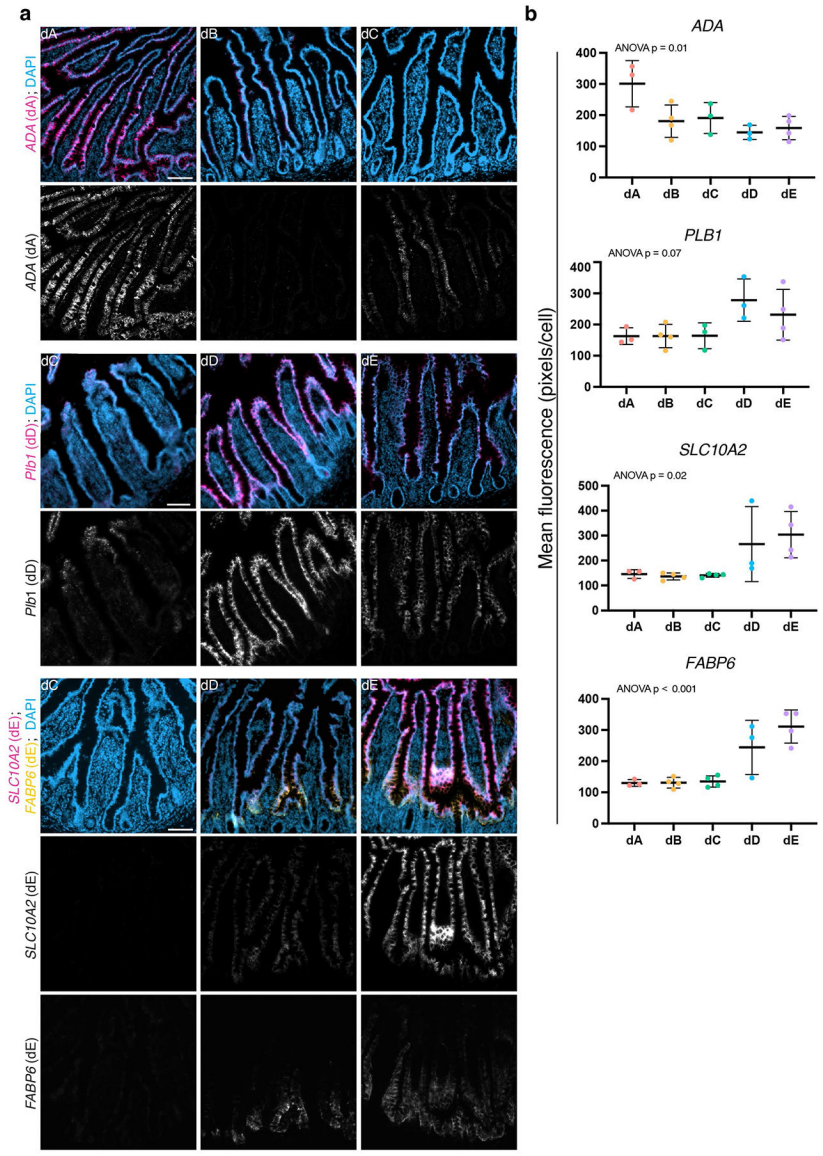
a *Right:* Average expression of the top 150 upregulated genes in enterocytes from human donor 1 in each segment, with segment order and hierarchical clustering based on expression distance between segments. Vertical white lines show the five domains that divide the small intestine, based on (*center*) gap statistics for hierarchical clusters of enterocytes in regional gene expression distance. Data bars are presented as mean values +/− confidence interval, based on all cells within the sample. *Right:* Cuts of dendrogram with optimal cluster number (magenta bracket, center). **b** Most highly regionalized genes expressed by enterocytes in mouse and donor 2 as in Fig. 1f,g but with a smaller number of genes displayed (75–100), as indicated on the y-axis. **c** Jensen-Shannon Divergence between enterocytes from segment pairs across the intestine of each individual mouse, with segment pair order and hierarchical clustering based on divergence values between segments. **d** Murine villus height by domain, presented as mean values +/− standard error of mean. Villus base to tip distances were

measured for 3–5 villi in each segment, for each of 4 mice. Statistical significance was calculated using one-way ANOVA followed by Tukey's multiple comparisons test for villus heights across all segments in each domain. *P < 0.05, ****P < 0.0001, ns not significant. **e** Domain-defining gene expression scores for human donor 1, as in Fig. 2c,d, coloured by domain with surrounding grey standard error bounds, across intestinal segments. Positions of domain boundaries calculated in **b** are noted with dotted lines and brackets. **f** Expression of key domain marker genes in mouse enterocytes across segments. The segment positions of each domain designation are indicated (bottom).



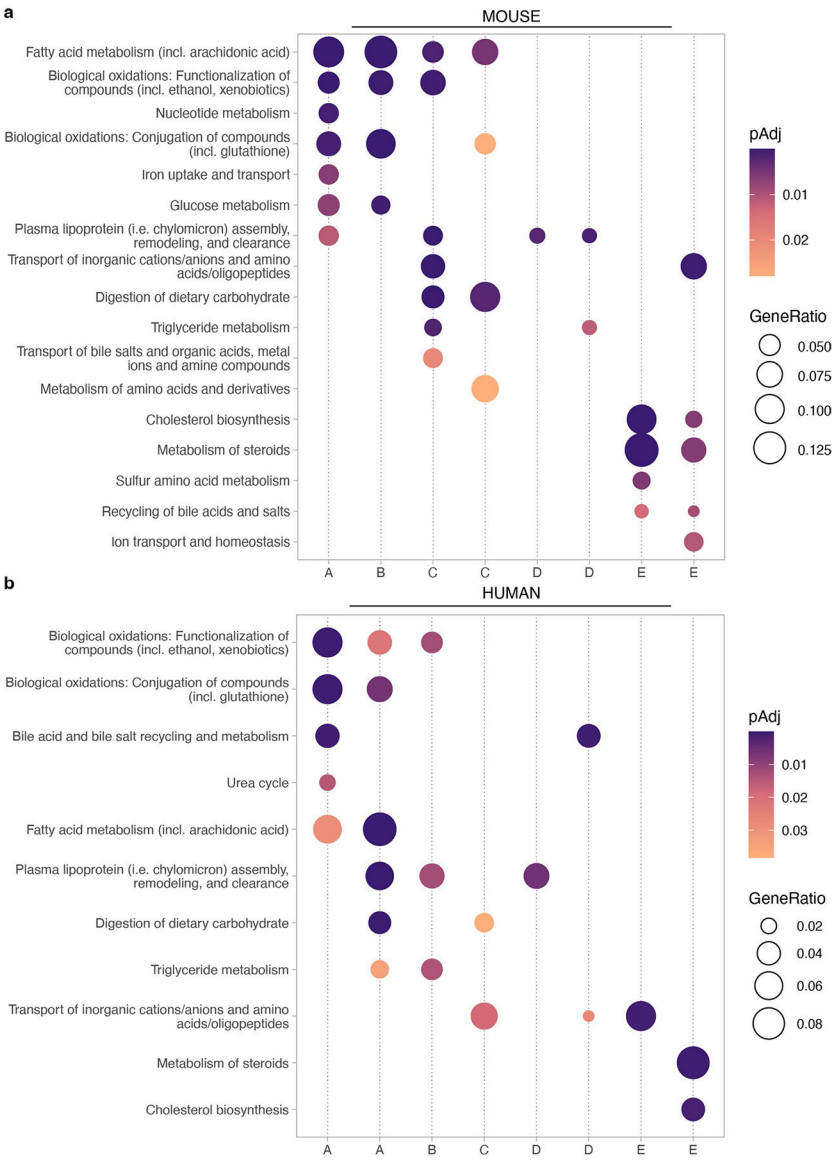
Extended Data Fig. 5 |. Single-molecule ISH validation of additional domain markers.
a,b Full-length murine intestinal tissue coiled from the proximal (outside) end to the distal (inside) end, probed with single-molecule ISH for select marker genes of domains as indicated. Channels are shown both individually and merged with pseudocolouring. White

boxes indicate insets. Scale bars are 2 mm, and 100 μm for insets. Similar results obtained with 3 mice.



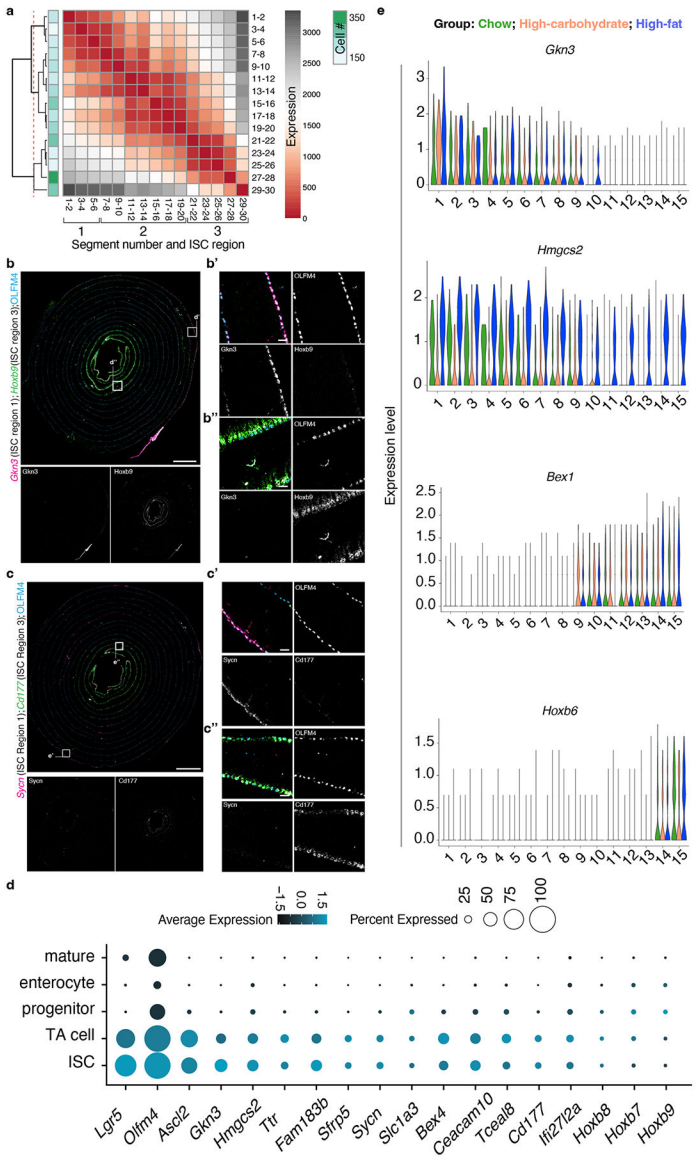
Extended Data Fig. 6 |. Domain marker expression in human tissue.

a Single channels of multi-channel images in Fig. 3b. Data are human tissue sections from indicated domains probed using single-molecule ISH with domain marker genes. Scale bars are 100 μm. **b** Quantification of mean fluorescence per domain for each donor, presented as mean values +/− standard error of mean. n = 3 or 4 donors per domain as indicated by number of datapoints. One-way ANOVA was performed to compare mean fluorescence in each donor by domain, p values for each marker are labeled.



Extended Data Fig. 7 |. Functional pathways enriched in domain-associated NMF gene modules in mouse and human.

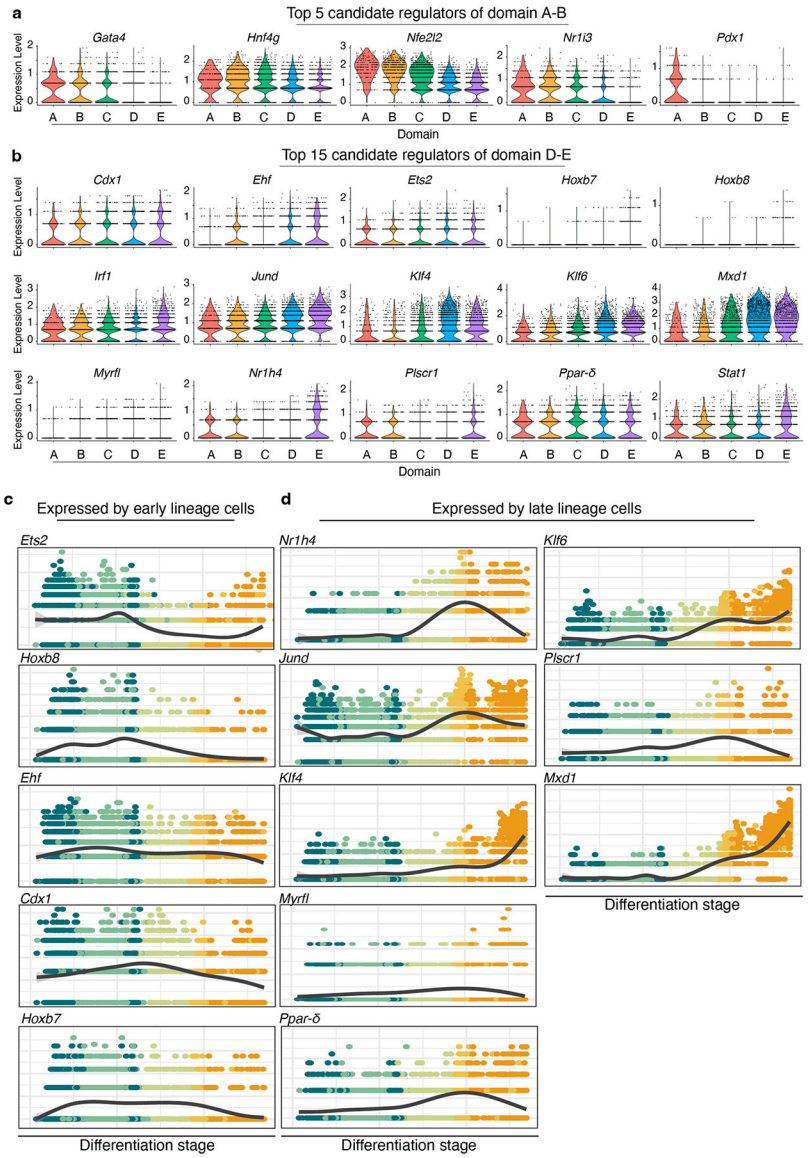
a,b Selected enriched functional pathways in each NMF gene module displayed in Fig. 2e,f in (**a**) mouse and (**b**) human. All gene modules with a regionally variable expression profile across segments that contained genes that encode aspects of nutrient metabolism are displayed (8 modules per species, dotted vertical lines). Module labels (bottom) are the domain(s) most closely-associated with each module, as determined by regional expression profile and rank of key domain-associated signature genes. Pathways were edited to remove redundancy. P values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure.



Extended Data Fig. 8 |. Divisions between regional intestinal stem cells (ISCs).

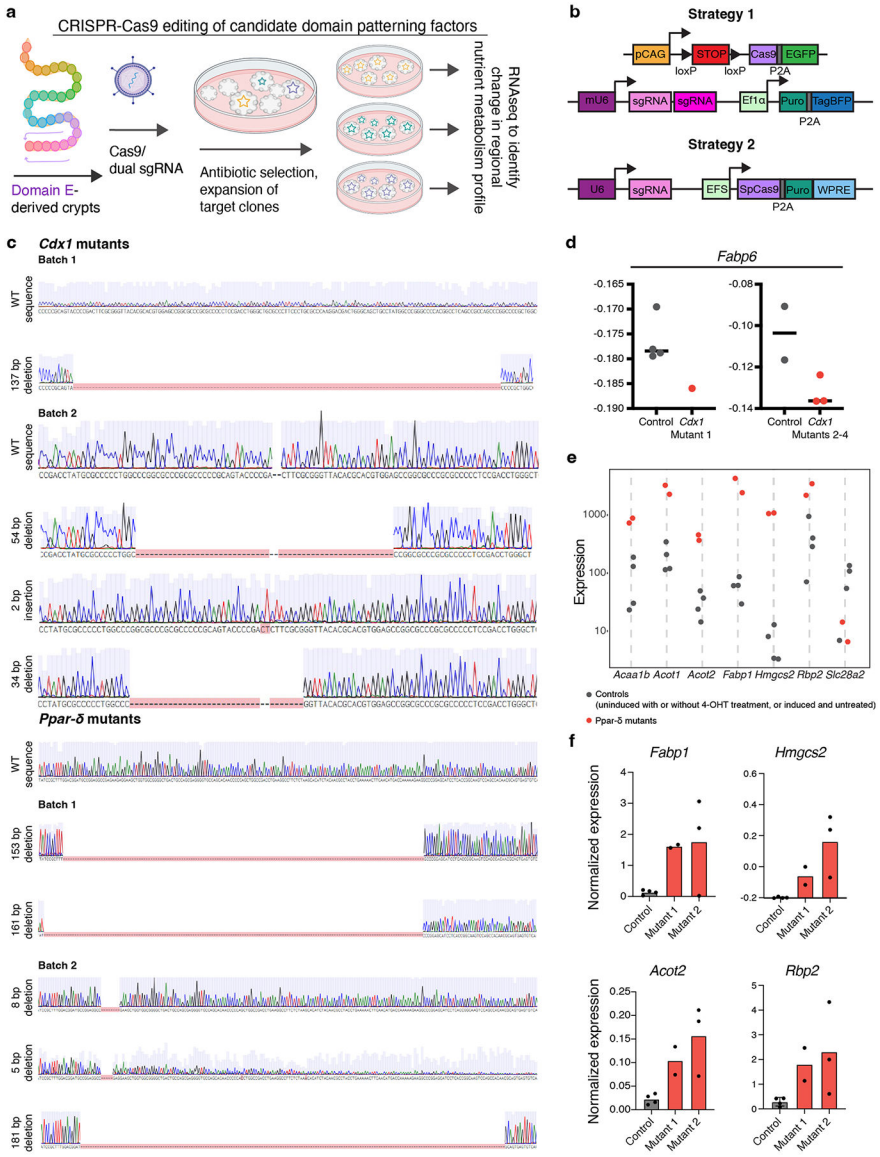
a Jensen-Shannon Divergence between ISCs from segment pairs across the intestine, with segment pair order and hierarchical clustering based on divergence values between segments. Dotted red line indicates level of hierarchical tree of domain divisions. **b,c** Full-length murine intestinal tissue coiled from the proximal (outside) end to the distal (inside) end, probed with single-molecule ISH for select regional ISC marker genes (as in Fig. 5d) as indicated. Channels are shown both individually and merged with pseudocolouring. White boxes indicate insets. Scale bars are 2 mm, and 100 μm for insets. Similar results obtained with 3 mice. **d** Expression of regional ISC marker genes in absorptive lineage cells. Dot colour reflects average expression, dot size reflects the percent of cells of each type expressing the marker. 'Mature' and 'progenitor' refer to enterocyte state. **e** Expression of ISC region 1 genes (*Gkn3* and *Hmgcs2*) and ISC region 3 genes (*Bex1* and *Hoxb6*) across ISCs from 15 segments collected from the small intestines of mice fed chow, high-carbohydrate,

or high-fat diets as indicated by colour. (n = 3 mice per diet group). ISC, intestinal stem cell; TA, transit amplifying.



Extended Data Fig. 9 |. Top candidate regulators of domain identity.

a,b Domain-wise expression levels of 5 candidate regulators of domain A and B identities (**a**) and 15 candidate regulators of domain D and E identities (**b**), identified using ChEA3 and SCENIC analyses. **c,d** Expression trajectories of indicated factors, coloured according to inferred differentiation stage in Fig. 6c. Transcription factor expression trajectories were plotted for cells in domain E. Plots are grouped according to expression by early-lineage cells (**c**) or differentiated cells (**d**).



Extended Data Fig. 10 |. Generation and analysis of *Ppar-δ* and *Cdx1* mutant domain E organoids.

a, b Schematics of CRISPR/Cas9 gene targeting strategy. Cas9 endonuclease was encoded in an endogenous genomic locus and 4-hydroxytamoxifen-induced (strategy 1) or delivered by lentiviral vector (strategy 2). Target-specific sgRNAs were delivered by lentiviral vectors (strategies 1 and 2) to induce mutations in the protein coding regions of the target genes. Following mutagenesis, selected clones were expanded and genotyped. Clones containing exclusively deleterious alleles were used for downstream analysis. **c** *Cdx1* mutant organoid sequences from CRISPR editing strategy 1 (‘batch 1’, n = 1 mutant line from mouse 1) and 2 (‘batch 2’, n = 3 unique mutant lines from mouse 2), and *Ppar-δ* mutant organoid sequences from editing strategy 1 (‘batch 1’, n = 2 unique mutant lines from mouse 1) and 2 (‘batch 2’, n = 3 unique mutant lines from mouse 2). Indel mutations are specified. **d** Trend towards decreased expression of *Fabp6* in Cdx1 mutant lines in both batches of mRNAseq expression data from editing strategies 1 and 2, which could not be merged.

Line represents median. **e** Expression of differentially expressed genes in individual *Ppar-δ* mutant organoid lines from batch 1 mutants (red dots) and control organoid lines (black dots). Batch 2 expression data of these and other DEGs in Fig. 6f,h. **f** Normalized mRNA levels of select DEGs of interest in *Ppar-δ* mutant organoids, validated with real time PCR. (n = 2–4 technical replicates per one control and two mutant organoid lines as indicated). bp, base pair; DEGs, differentially expressed genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to H. Miyazaki, D. Castillo-Azofeifa and other members of the Klein laboratory for valuable discussions, experimental assistance and protocol development. We thank M. Helmrath, N. Shroyer, Y.-H. Lo and members of the Intestinal Stem Cell Consortium for critical scientific input throughout this project. We also thank B. Ohlstein, I. Chen, K. Bahar Halpern, Z. Sullivan, D. Conrad, J. Sheu-Gruttadauria, J. Bush, and E. Chow for sharing data, resources and expertise. This study benefited from the following cores and facilities at UCSF: the Center for Advanced Technology, the Institute for Human Genetics, Parnassus Flow Cytometry Core, ViraCore, VIable Tissue Acquisition Lab, the Biological Imaging Development CoLab, and the Laboratory Animal Resource Center. Portions of schematic figure panels were created with BioRender. com. This work was funded by NIH R35-DE026602 and U01DK103147 from the Intestinal Stem Cell Consortium, a collaborative research project funded by the National Institute of Diabetes and Digestive and Kidney Diseases and the National Institute of Allergy and Infectious Diseases (to O.D.K.). R.K.Z. was supported by NIH F32 DK125089 and an American Cancer Society—South Florida Research Council Postdoctoral Fellowship (PF-20-037-01-DDC). Finally, our most sincere gratitude goes to Donor Network West, and to the organ donors and their families for their generosity in supporting basic science research.

Data availability

The datasets generated and analysed in the current study are available in the Gene Expression Omnibus (GEO; BioProject accession code GSE201859) and can be visualized using the Chan Zuckerberg CELLxGENE tool (https://cellxgene.cziscience.com/collections/3db5617e-9f12-4eb4-8416-94893a0d7c46). Previously published single-cell sequencing data[6] analysed here are available under accession code GSE92332. Source data and analysis outputs are provided in the Supplementary Information associated with this publication. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Code availability

Custom code developed for this manuscript is available in Zenodo under record no. 10223562.

References

1. San Roman AK & Shivdasani RA Boundaries, junctions and transitions in the gastrointestinal tract. Exp. Cell. Res 317, 2711–2718 (2011). [PubMed: 21802415]
2. Brown H & Esterhazy D Intestinal immune compartmentalization: implications of tissue specific determinants in health and disease. Mucosal Immunol. 14, 1259–1270 (2021). [PubMed: 34211125]
3. Esterhazy D. et al. Compartmentalized gut lymph node drainage dictates adaptive immune responses. Nature 569, 126–130 (2019). [PubMed: 30988509]

4. Altmann GG & Leblond CP Factors influencing villus size in the small intestine of adult rats as revealed by transposition of intestinal segments. Am. J. Anat 127, 15–36 (1970). [PubMed: 5412637]
5. Bates MD et al. Novel genes and functional relationships in the adult mouse gastrointestinal tract identified by microarray analysis. Gastroenterology 122, 1467–1482 (2002). [PubMed: 11984531]
6. Haber AL et al. A single-cell survey of the small intestinal epithelium. Nature 551, 333–339 (2017). [PubMed: 29144463]
7. Elmentaite R. et al. Cells of the human intestinal tract mapped across space and time. Nature 597, 250–255 (2021). [PubMed: 34497389]
8. Burclaff J. et al. A proximal-to-distal survey of healthy adult human small intestine and colon epithelium by single-cell transcriptomics. Cell Mol. Gastroenterol. Hepatol 13, 1554–1589 (2022). [PubMed: 35176508]
9. Wang Y. et al. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. J. Exp. Med 217, jem.20191130 (2020).
10. Hickey JW et al. Organization of the human intestine at single-cell resolution. Nature 619, 572–584 (2023). [PubMed: 37468586]
11. Fawkner-Corbett D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. Cell 184, 810–826 (2021). [PubMed: 33406409]
12. Zwick RK, Ohlstein B & Klein OD Intestinal renewal across the animal kingdom: comparing stem cell activity in mouse and *Drosophila*. Am. J. Physiol. Gastrointest. Liver Physiol 316, G313–G322 (2019). [PubMed: 30543448]
13. Buchon N. et al. Morphological and molecular characterization of adult midgut compartmentalization in *Drosophila*. Cell Rep. 3, 1725–1738 (2013). [PubMed: 23643535]
14. Marianes A & Spradling AC Physiological and stem cell compartmentalization within the *Drosophila* midgut. eLife 2, e00886 (2013). [PubMed: 23991285]
15. Driver I & Ohlstein B Specification of regional intestinal stem cell identity during *Drosophila* metamorphosis. Development 141, 1848–1856 (2014). [PubMed: 24700821]
16. Hudry B. et al. Sex differences in intestinal carbohydrate metabolism promote food intake and sperm maturation. Cell 178, 901–918 (2019). [PubMed: 31398343]
17. Middendorp S. et al. Adult stem cells in the small intestine are intrinsically programmed with their location-specific function. Stem Cells 32, 1083–1091 (2014). [PubMed: 24496776]
18. Kayisoglu O. et al. Location-specific cell identity rather than exposure to GI microbiota defines many innate immune signalling cascades in the gut epithelium. Gut 70, 687–697 (2021). [PubMed: 32571970]
19. Kraiczy J. et al. DNA methylation defines regional identity of human intestinal epithelial organoids and undergoes dynamic changes during development. Gut 68, 49–61 (2019). [PubMed: 29141958]
20. McGinnis CS et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods 16, 619–626 (2019). [PubMed: 31209384]
21. Moor AE et al. Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. Cell 175, 1156–1167 (2018). [PubMed: 30270040]
22. Tibshirani R, Walther G & Hastie T Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B Stat. Methodol 63, 411–423 (2001).
23. Peng M, Li Y, Wamsley B, Wei Y & Roeder K Integration and transfer learning of single-cell transcriptomes via cFIT. Proc. Natl Acad. Sci. USA 118, e2024383118 (2021). [PubMed: 33658382]
24. Sullivan ZA et al. γδ T cells regulate the intestinal response to nutrient sensing. Science 371, eaba8310 (2021). [PubMed: 33737460]
25. Enriquez JR et al. A dietary change to a high-fat diet initiates a rapid adaptation of the intestine. Cell Rep. 41, 111641 (2022). [PubMed: 36384107]
26. Goda T. Regulation of the expression of carbohydrate digestion/absorption-related genes. Br. J. Nutr 84, S245–S248 (2000). [PubMed: 11242478]
27. Clara R. et al. Metabolic adaptation of the small intestine to short- and medium-term high-fat diet exposure. J. Cell. Physiol 232, 167–175 (2017). [PubMed: 27061934]

28. Ko C-W, Qu J, Black DD & Tso P Regulation of intestinal lipid metabolism: current concepts and relevance to disease. Nat. Rev. Gastroenterol. Hepatol 17, 169–183 (2020). [PubMed: 32015520]
29. Gebert N. et al. Region-specific proteome changes of the intestinal epithelium during aging and dietary restriction. Cell Rep. 31, 107565 (2020). [PubMed: 32348758]
30. Biton M. et al. T helper cell cytokines modulate intestinal stem cell renewal and differentiation. Cell 175, 1307–1320 (2018). [PubMed: 30392957]
31. Maimets M. et al. Mesenchymal-epithelial crosstalk shapes intestinal regionalisation via Wnt and Shh signalling. Nat. Commun 13, 715 (2022). [PubMed: 35132078]
32. Spence JR, Lauf R & Shroyer NF Vertebrate intestinal endoderm development. Dev. Dyn 240, 501–520 (2011). [PubMed: 21246663]
33. Thompson CA, DeLaForest A & Battle MA Patterning the gastrointestinal epithelium to confer regional-specific functions. Dev. Biol 435, 97–108 (2018). [PubMed: 29339095]
34. Thompson CA et al. GATA4 is sufficient to establish jejunal versus ileal identity in the small intestine. Cell Mol. Gastroenterol 3, 422–446 (2017).
35. Chen C, Fang RX, Davis C, Maravelias C & Sibley E Pdx1 inactivation restricted to the intestinal epithelium in mice alters duodenal gene expression in enterocytes and enteroendocrine cells. Am. J. Physiol. Gastrintest. Liver Physiol 297, G1126–G1137 (2009).
36. Battle MA et al. GATA4 is essential for jejunal function in mice. Gastroenterology 135, 1676–1686 (2008). [PubMed: 18812176]
37. Bosse T. et al. Gata4 is essential for the maintenance of jejunal-ileal identities in the adult mouse small intestine. Mol. Cell. Biol 26, 9060–9070 (2006). [PubMed: 16940177]
38. Keenan AB et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 47, W212–W224 (2019). [PubMed: 31114921]
39. Aibar S. et al. SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 14, 1083–1086 (2017). [PubMed: 28991892]
40. Zorn AM & Wells JM Vertebrate endoderm development and organ formation. Annu. Rev. Cell Dev. Biol 25, 221–251 (2009). [PubMed: 19575677]
41. Verzi MP, Shin H, Ho LL, Liu XS & Shivdasani RA Essential and redundant functions of caudal family proteins in activating adult intestinal genes. Mol. Cell. Biol 31, 2026–2039 (2011). [PubMed: 21402776]
42. Hryniuk A, Grainger S, Savory JGA & Lohnes D Cdx function is required for maintenance of intestinal identity in the adult. Dev. Biol 363, 426–437 (2012). [PubMed: 22285812]
43. Bonhomme C. et al. Cdx1, a dispensable homeobox gene for gut development with limited effect in intestinal cancer. Oncogene 27, 4497–4502 (2008). [PubMed: 18372917]
44. Doktorova M. et al. Intestinal PPARdelta protects against diet-induced obesity, insulin resistance and dyslipidemia. Sci. Rep 7, 846 (2017). [PubMed: 28404991]
45. Beyaz S. et al. High-fat diet enhances stemness and tumorigenicity of intestinal progenitors. Nature 531, 53–58 (2016). [PubMed: 26935695]
46. Mana MD et al. High-fat diet-activated fatty acid oxidation mediates intestinal stemness and tumorigenicity. Cell Rep. 35, 109212 (2021). [PubMed: 34107251]
47. Seiler KM et al. Single-cell analysis reveals regional reprogramming during adaptation to massive small bowel resection in mice. Cell Mol. Gastroenterol. Hepatol 8, 407–426 (2019). [PubMed: 31195149]
48. Nusse YM et al. Parasitic helminths induce fetal-like reversion in the intestinal stem cell niche. Nature 559, 109–113 (2018). [PubMed: 29950724]
49. Schneider C. et al. A metabolite-triggered tuft cell-ILC2 circuit drives small intestinal remodeling. Cell 174, 271–284 (2018). [PubMed: 29887373]
50. Cheng CW et al. Ketone body signaling mediates intestinal stem cell homeostasis and adaptation to diet. Cell 178, 1115–1131 (2019). [PubMed: 31442404]
51. Stine RR et al. PRDM16 maintains homeostasis of the intestinal epithelium by controlling region-specific metabolism. Cell Stem Cell 25, 830–845 (2019). [PubMed: 31564549]

52. Obniski R, Sieber M & Spradling AC Dietary lipids modulate notch signaling and influence adult intestinal development and metabolism in *Drosophila*. Dev. Cell 47, 98–111 (2018). [PubMed: 30220569]
53. Gajendran M, Loganathan P, Catinella AP & Hashash JG A comprehensive review and update on Crohn's disease. Dis. Mon 64, 20–57 (2018). [PubMed: 28826742]
54. Pan SY & Morrison H Epidemiology of cancer of the small intestine. World J. Gastrointest. Oncol 3, 33–42 (2011). [PubMed: 21461167]
55. Schottenfeld D, Beebe-Dimmer JL & Vigneau FD The epidemiology and pathogenesis of neoplasia in the small intestine. Ann. Epidemiol 19, 58–69 (2009). [PubMed: 19064190]
56. Brown EM, Clardy J & Xavier RJ Gut microbiome lipid metabolism and its impact on host physiology. Cell Host Microbe 31, 173–186 (2023). [PubMed: 36758518]
57. Tian H. et al. A reserve stem cell population in small intestine renders *Lgr5*-positive cells dispensable. Nature 478, 255–259 (2011). [PubMed: 21927002]
58. Huch M. et al. In vitro expansion of single $Lgr5^+$ liver stem cells induced by Wnt-driven regeneration. Nature 494, 247–250 (2013). [PubMed: 23354049]
59. Platt RJ et al. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell 159, 440–455 (2014). [PubMed: 25263330]
60. Madisen L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. Nat. Neurosci 13, 133–140 (2010). [PubMed: 20023653]
61. Smillie CS et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. Cell 178, 714–730 (2019). [PubMed: 31348891]
62. Melsted P. et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nat. Biotechnol 39, 813–818 (2021). [PubMed: 33795888]
63. Lun ATL et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 20, 63 (2019). [PubMed: 30902100]
64. Zhu Q, Conrad DN & Gartner ZJ deMULTIplex2: Robust Sample Demultiplexing for scRNA-seq (Cold Spring Harbor Laboratory, 2023).
65. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q & Powell JE *scPred*: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. 20, 264 (2019). [PubMed: 31829268]
66. Stuart T. et al. Comprehensive integration of single-cell data. Cell 177, 1888–1902 (2019). [PubMed: 31178118]
67. Hao Y. et al. Integrated analysis of multimodal single-cell data. Cell 184, 3573–3587 (2021). [PubMed: 34062119]
68. Hafemeister C & Satija R Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 20, 296 (2019). [PubMed: 31870423]
69. Cunningham F. et al. Ensembl 2022. Nucleic Acids Res. 50, D988–D995 (2022). [PubMed: 34791404]
70. Kolde R. Pheatmap: pretty heatmaps. R package version 1.2 (CRAN.R Project, 2012); https://cran.r-project.org/package=pheatmap
71. Lin JH Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory 37, 145–151 (1991).
72. Drost H-G Philentropy: information theory and distance quantification with R.J. Open Source Softw 3, 765 (2018).
73. Kotliar D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. eLife 8, e43803 (2019). [PubMed: 31282856]
74. Wu TZ et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation 2, 100141 (2021). [PubMed: 34557778]
75. van de Sande B. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat. Protoc 15, 2247–2276 (2020). [PubMed: 32561888]
76. Street K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics 19, 477 (2018). [PubMed: 29914354]

77. Schindelin J. et al. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9, 676–682 (2012). [PubMed: 22743772]

78. Bankhead P. et al. QuPath: Open source software for digital pathology image analysis. Sci. Rep 7, 16878 (2017). [PubMed: 29203879]

79. Castillo-Azofeifa D. et al. Atoh1^{+} secretory progenitors possess renewal capacity independent of Lgr5^{+} cells during colonic regeneration. EMBO J. 38, e99984 (2019). [PubMed: 30635334]

80. McKinley KL Employing CRISPR/Cas9 genome engineering to dissect the molecular requirements for mitosis. Methods Cell. Biol 144, 75–105 (2018). [PubMed: 29804684]

81. Gilbert LA et al. Genome-scale CRISPR-mediated control of gene repression and activation. Cell 159, 647–661 (2014). [PubMed: 25307932]

82. Adamson B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell 167, 1867–1882 (2016). [PubMed: 27984733]

83. Sanjana NE, Shalem O & Zhang F Improved vectors and genome-wide libraries for CRISPR screening. Nat. Methods 11, 783–784 (2014). [PubMed: 25075903]

84. Koo BK et al. Controlled gene expression in primary Lgr5 organoid cultures. Nat. Methods 9, 81–83 (2012).

85. Bray NL, Pimentel H, Melsted P & Pachter L Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol 34, 525–527 (2016). [PubMed: 27043002]

86. Risso D, Ngai J, Speed TP & Dudoit S Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol 32, 896–902 (2014). [PubMed: 25150836]

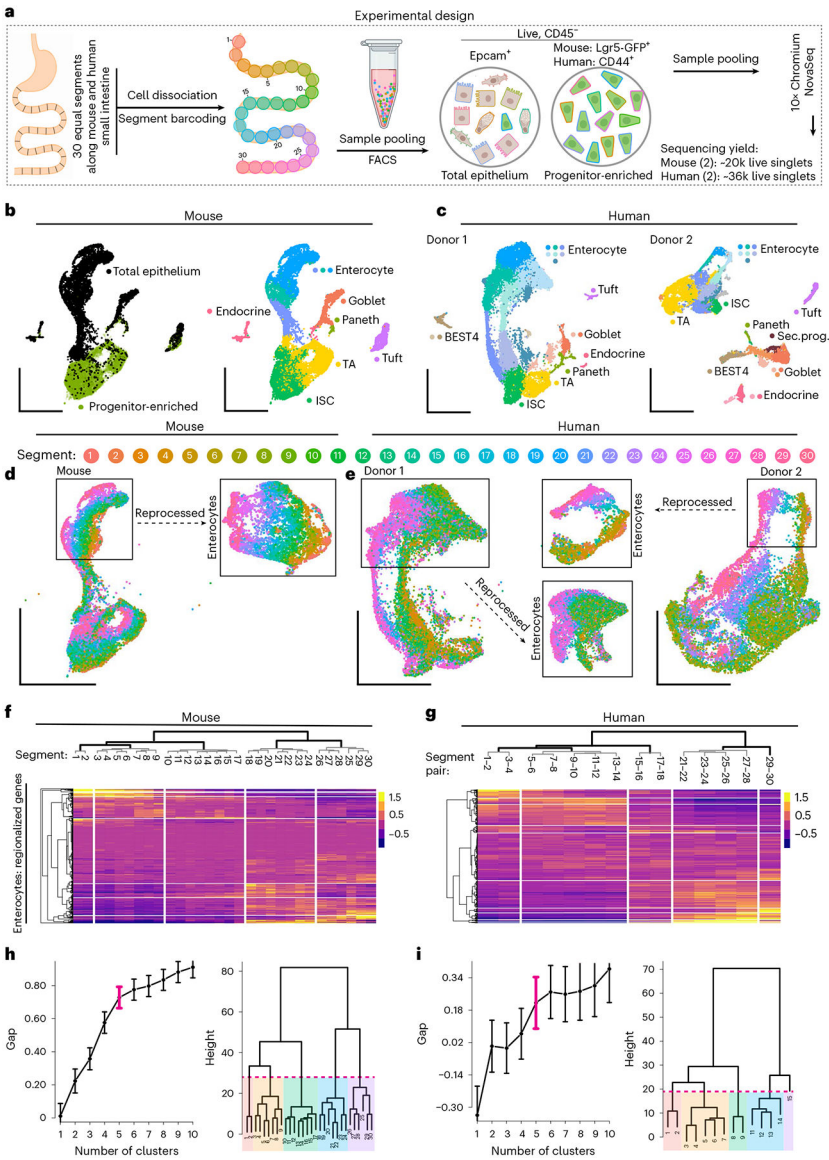


Fig. 1 |. Five enterocyte groups occupy distinct zones along the length of the SI.

a, scRNA-seq of epithelial cells from 30 equal segments of the mouse ($n = 2$) and human ($n = 2$) SI. Cells from each segment were dissociated, tagged with segment-specific barcodes, pooled, sorted into total epithelial and progenitor-enriched samples, and sequenced. Cell number yields following data quality control (QC) are shown. **b**,**c**, Uniform manifold approximation and projection (UMAP) of sequenced mouse (**b**) and human (**c**) cells following QC, annotated with sample identification (**b**, left) or predicted cell type. Microfold (M-) cells not displayed; c.f. Extended Data Figs. 2 and 3a. **d**,**e**, UMAP of absorptive cells coloured by lengthwise segment number. Insets display reprocessed enterocyte subsets. **f**–**i**, Average expression of the top 150 upregulated genes in mouse (**f**) and human (**g**) enterocytes in each segment, with segment order and hierarchical clustering based on expression distance between segments (**h**,**i**). Vertical white lines in **f** and **g** show domain delineations, based on **h** and **i**, respectively. Left (**h**,**i**): gap statistics for hierarchical clusters

of enterocytes in regional gene expression distance. Data bars are presented as mean values ± confidence interval, based on all cells within the sample. Right (**h**,**i**): cuts of dendrograms with optimal cluster numbers (magenta brackets, left). The five resulting regional enterocyte groups are shaded.

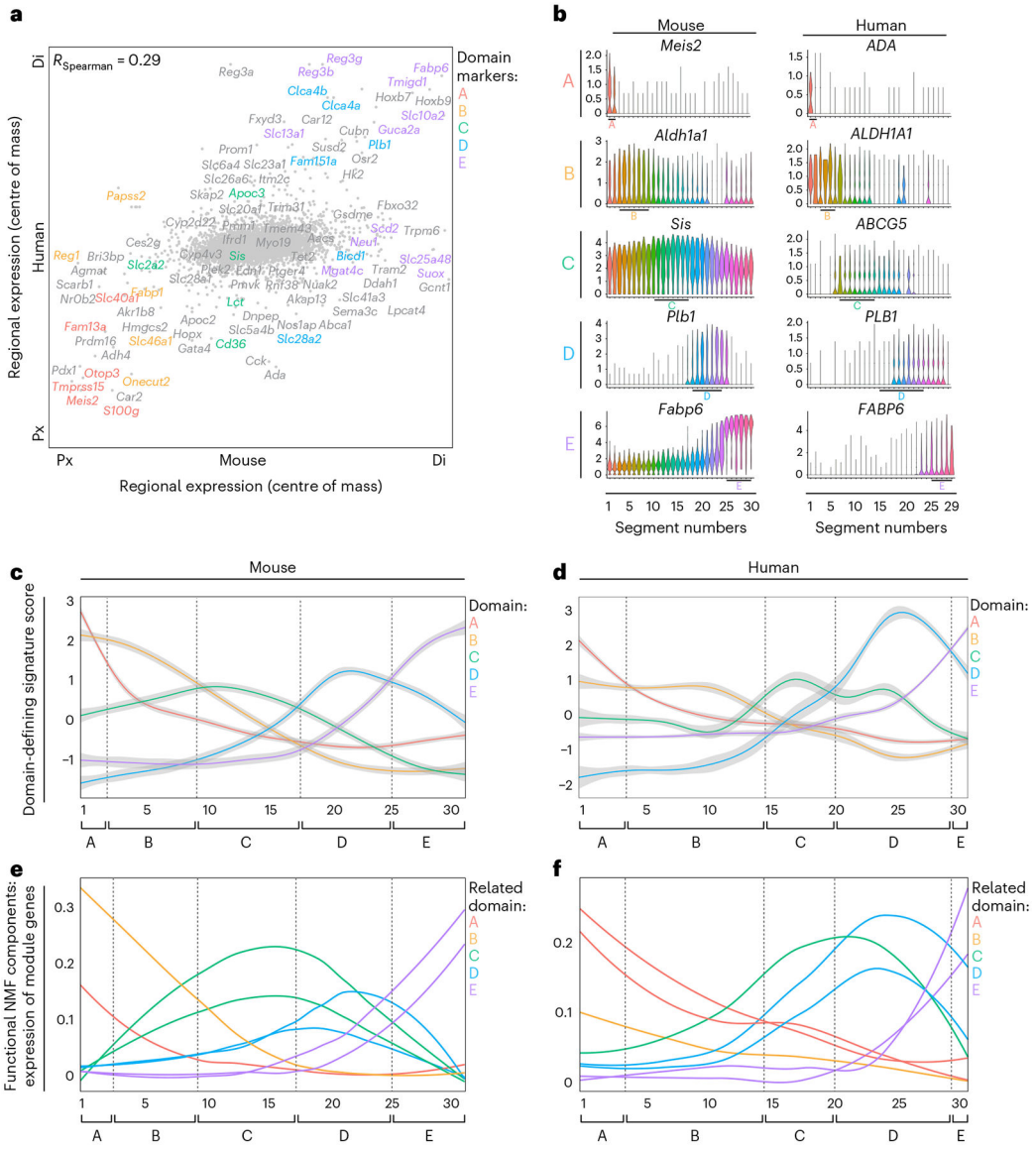


Fig. 2 |. A progression of five distinct gene modules divides intestinal length.

a, Comparison of segment centres of mass for 6,191 homologous genes in mouse and human enterocytes, with mean sum-normalized levels of >1 × 10−5 in at least one point along intestinal length in both species. $R_{Spearman}$ = 0.29, n = 2 mice and 2 human donors. The top segmentally variable genes in each species are shown, with mouse domain signature genes colour-coded as indicated. Px and Di identify the proximal and distal ends of the mouse (x axis) and human (y axis) SI. **b**, Expression level by segment of select marker genes of each domain in mouse and human enterocytes. Human genes were domain-enriched in both donors, and representative plots from donor 1 are shown. **c**,**d**, Domain-defining gene expression scores for mouse (**c**) and human donor 2 (**d**), which represent the mean scaled expression of the top 20 domain-defining genes, coloured by domain, with surrounding grey standard error bounds, across intestinal segments. Segment positions are numbered (x axis), and the positions of domain boundaries calculated in Fig. 1h,i are noted with

dotted lines and brackets. **e**,**f**, Cumulative expression of regionally variable mouse (**e**) and human donor 2 (**f**) NMF gene modules across intestinal segments. Gene modules that encode physiological functions associated with nutrient metabolism are displayed. Module lines are coloured according to the domain A–E they most closely resemble based on regional expression trajectory and signature gene expression. Segment positions are numbered (*x* axis), and the positions of the domain boundaries calculated in Fig. 1h,i are noted with dotted lines and brackets. NMF, non-negative matrix factorization.

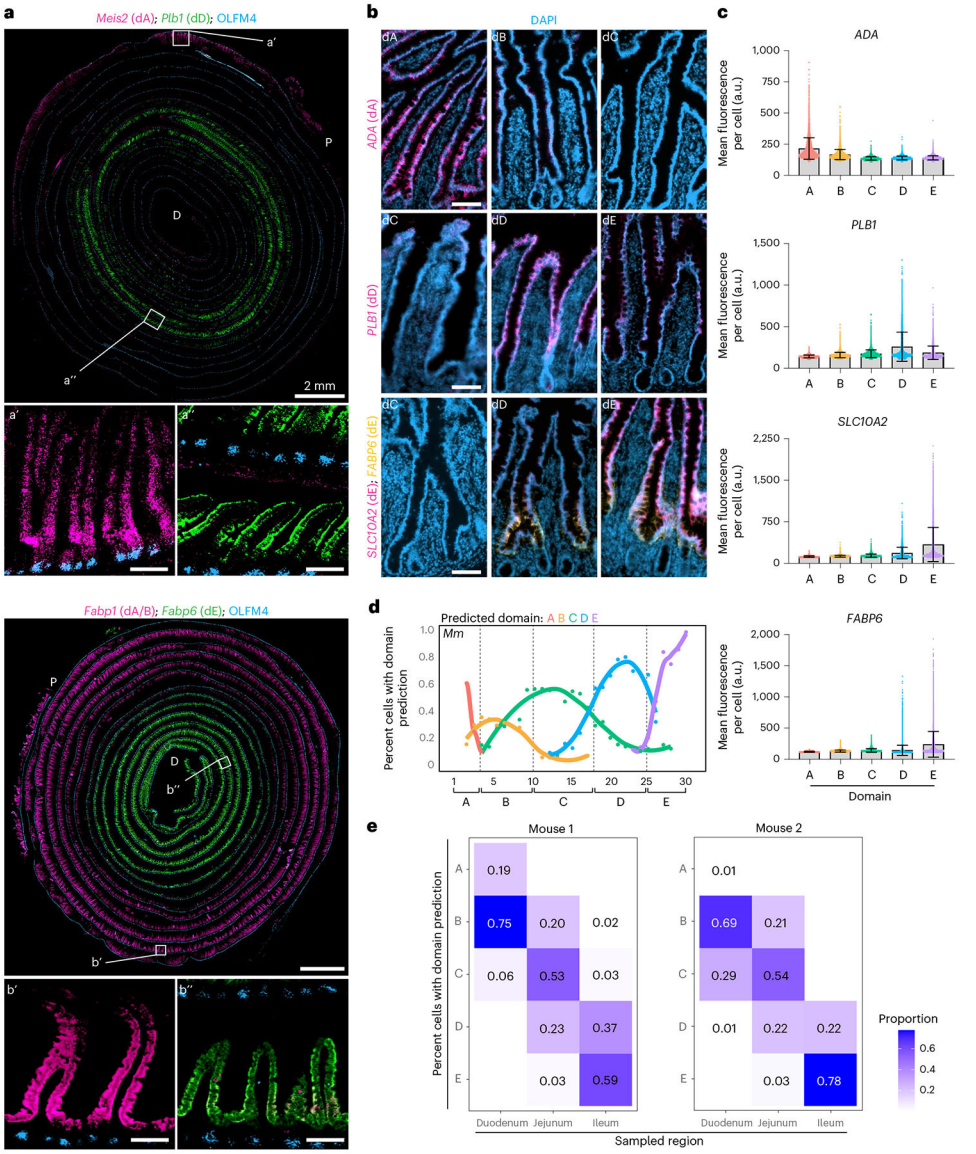


Fig. 3 |. Domain identity can be detected across samples and used for systematic classification of intestinal regions.

a, Full-length murine intestinal tissue coiled from the proximal (P, outside) to distal (D, inside) end, probed with single-molecule multiplexed ISH for select domain marker genes. White boxes mark the insets. Scale bars, 2 mm (main) and 100 μm (insets). Similar results were obtained with three mice**. b**,**c**, Images (**b**) of human tissue sections from the indicated domains, probed as in **a** for the indicated domain marker genes, and quantification (**c**) of the mean fluorescence per cell. Representative images and quantification from one donor are displayed. Similar results were obtained from four total donors. dA to dE indicate domains A to E. Scale bars, 100 μm. **d**,**e**, Predicted domain identities of enterocytes sequenced in mouse sequencing set two (test dataset, $n = 2$ mice; (**d**) and cells previously sequenced in published data[6] (**e**), as assigned by computational transfer of domain labels from the training dataset. In **d**, the proportion of cells with the domain predictions at each segment position (x axis) is indicated by line colour, and the dotted vertical lines indicate domain boundaries

in the training set in Fig. 1h. In **e**, the proportions of cells in the reported classic intestinal regions are indicated in each column. *Mm*, mouse; a.u., arbitrary units.

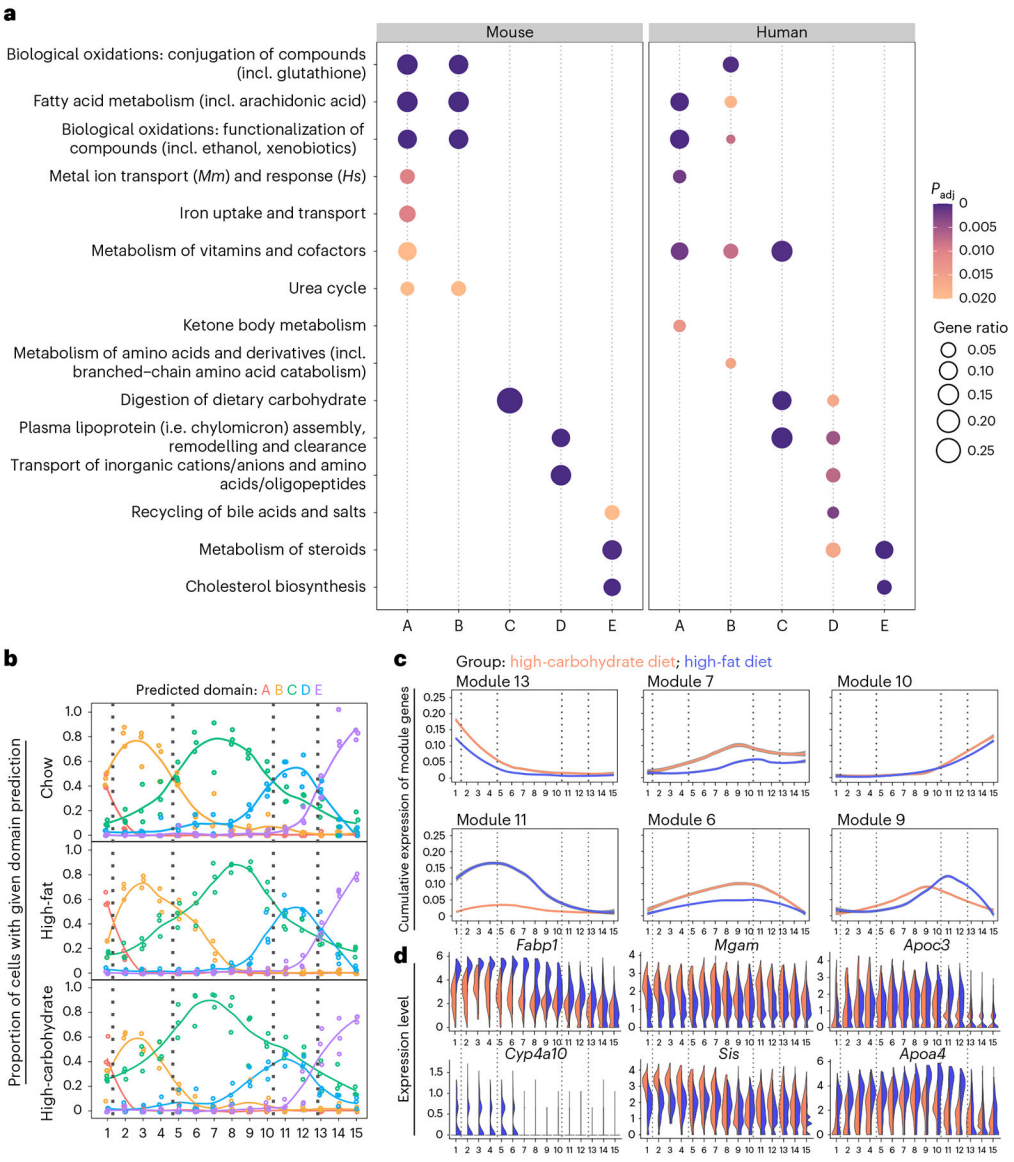


Fig. 4 |. Domains are associated with distinct aspects of nutrient metabolism.

a, Summary of pathway enrichment in each mouse and human domain, represented as circles coloured according to adjusted *P* values and sized according to gene ratio (ratio of domain marker genes that are annotated with the pathway term). Selected domain-enriched, nutrient metabolism-associated pathways with adjusted $P < 0.02$ are shown. *P* values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure. **b**, Predicted domain identities of sequenced enterocytes from mice administered a high-fat or high-carbohydrate diet for seven days ($n = 3$ mice per diet group), as assigned by computational transfer of domain labels from the mouse training dataset. The proportions of cells with the domain predictions in three mice per diet group are indicated by the colour of the best fit lines. Dots are data points from each mouse. Dotted vertical lines indicate domain boundary positions predicted for the chow diet group (top). **c**, Cumulative expression of regionally variable NMF gene modules associated with nutrient metabolism across intestinal segments

in each diet group, indicated by line colour. 95% confidence intervals are indicated with grey bands. **d**, Expression levels of select genes from the indicated modules associated with lipid metabolism (modules 11 and 9) and carbohydrate absorption (module 6) in mice fed high-fat (purple) or high-carbohydrate (orange) diets. Similar results were obtained with three mice. *Mm*, mouse; *Hs*, human.

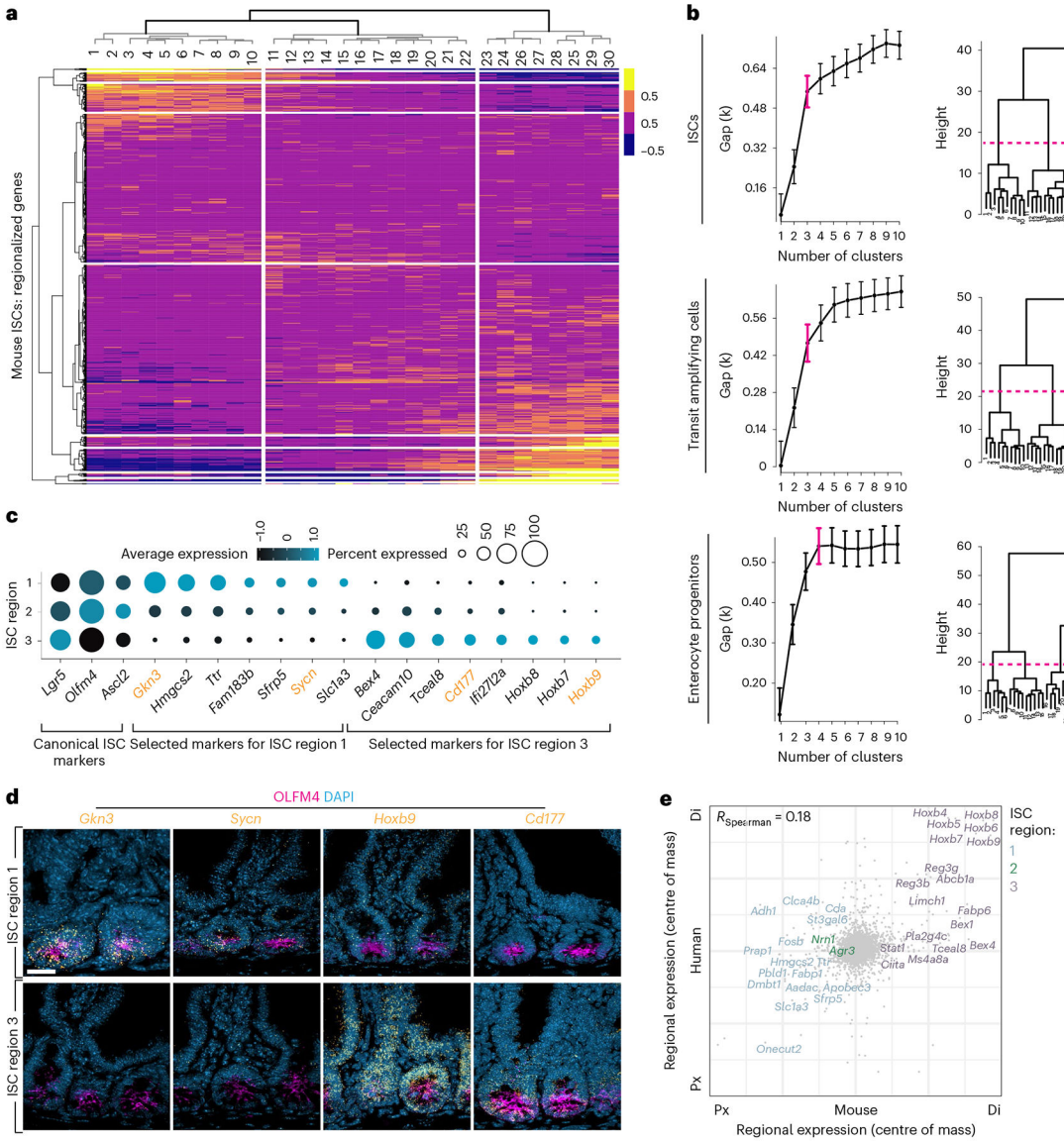


Fig. 5 |. Three regional stem cell populations reside within the SI.

a, Average expression of the top 100 upregulated genes in murine ISCs in each segment, with segment order and hierarchical clustering based on expression distance between segments. Vertical white lines mark the three domains that divide the ISC compartment, based on gap statistics**. b**, Left: gap statistics for clusters of regional gene expression in regional ISCs, transit amplifying cells and enterocyte progenitors. Right: cuts of dendrograms (dashed magenta lines) with optimal cluster numbers (magenta brackets, left) for each cell type. Data bars present mean values ± confidence interval, based on all cells within the sample. **c**, Selected regional ISC subpopulation marker genes, represented as dots coloured according to the average expression level and sized according to the percent of ISCs expressing the marker. Orange marker labels were validated with ISH (**d**). **d**, Intestinal crypts probed with single-molecule ISH for select regional ISC marker genes as indicated. Scale bars, 20 μm. **e**, Comparison of segment centres of mass for 7,668 homologous genes

in mouse and human crypt cells with mean sum-normalized levels >1 × 10^{-5} in at least one point along the intestinal length in both species. $R_{Spearman}$ = 0.18, n = 2 mice and two human donors. Top segmentally variable genes in each species are shown, and mouse regional ISC signature genes are colour-coded as indicated. Px and Di identify the proximal and distal ends of the mouse (x axis) and human (y axis) SI.

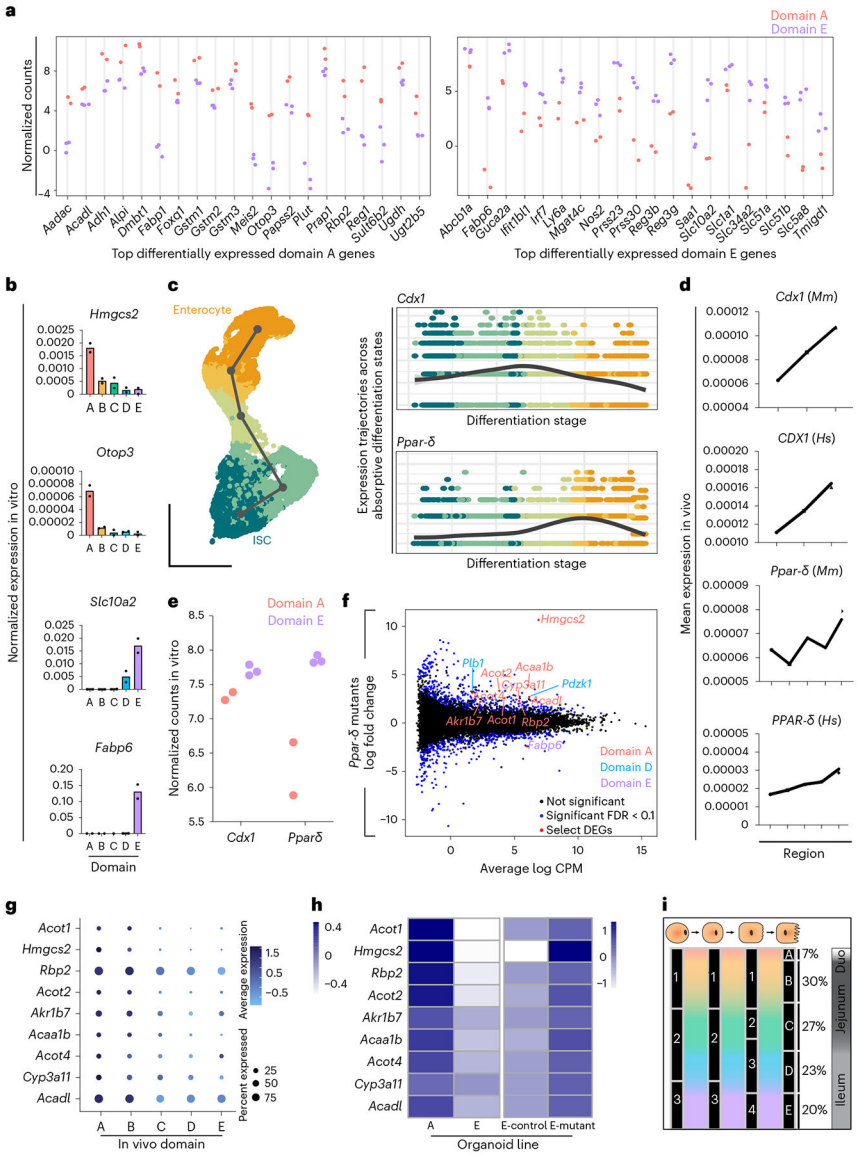


Fig. 6 |. Transcriptional control of enterocyte regional identity.

a, mRNA levels of the top 20 domain A (left) and domain E (right) signature genes most highly differentially expressed in domain A- (n = 2 lines) or E-derived organoids (n = 3 lines), respectively, evaluated with mRNA-seq. Organoid lines represent biological replicates and were assessed 5–6 days after passaging in long-term (>5 week) culture. **b**, qPCR confirmation of selected domain A (*Hmgcs2*, *Otop3*) and domain E (*Slc10a2*, *Fabp6*) signature genes in domain A- (n = 2 lines) and E-derived (n = 2 lines) organoids, respectively. **c**, UMAP of murine absorptive cells (left) and expression trajectories of *Cdx1* and *Ppar-δ* (right), in which all domain E cells are coloured according to inferred differentiation stage. **d**, Expression profiles of *Cdx1* in crypts across ISC regions (*Mm*) or equal thirds of intestinal length (*Hs*), and *Ppar-δ* in enterocytes across domains (*Mm*) or equal fifths of intestinal length (*Hs*). Data are presented as mean expression levels of cells in each position from mouse scRNA-seq data (as in Fig. 1a). **e**, mRNA levels of

Cdx1 and *Ppar-δ* in domain A- or E-derived organoids, as in **a**. **f**, Mean-difference plot of expression in *Ppar-δ* mutant organoids relative to controls. Dot colours are specified. Regionally variable differentially expressed genes that encode lipid metabolism are labelled and coloured by domain. $n = 3$ unique *Ppar-δ* mutant organoid lines and two control lines. **g**, Dotplot of in vivo expression levels of the domain A signature *Ppar-δ* mutant DEGs labelled in **f**. Dot size represents percent expressing enterocytes, and colour intensity represents average expression. DEG, differentially expressed gene. **h**, Heatmap showing mRNA levels of the domain A lipid metabolism signature in domain A- and E-derived organoids as in **a**, and in control and *Ppar-δ* knockout domain E organoids as in **f**. **i**, Summary of regional specialization of the SI. Within the absorptive lineage (schematized, top), we find three regional ISC populations, predicted to give rise to three TA cell populations, which produce four enterocyte progenitors that specialize into five mature enterocyte types that occupy absorption domains A–E. The estimated proportion of the intestinal length of each domain and our approximation of the corresponding traditional intestinal regions (gradient colours) are shown. *Mm*, mouse; *Hs*, human.